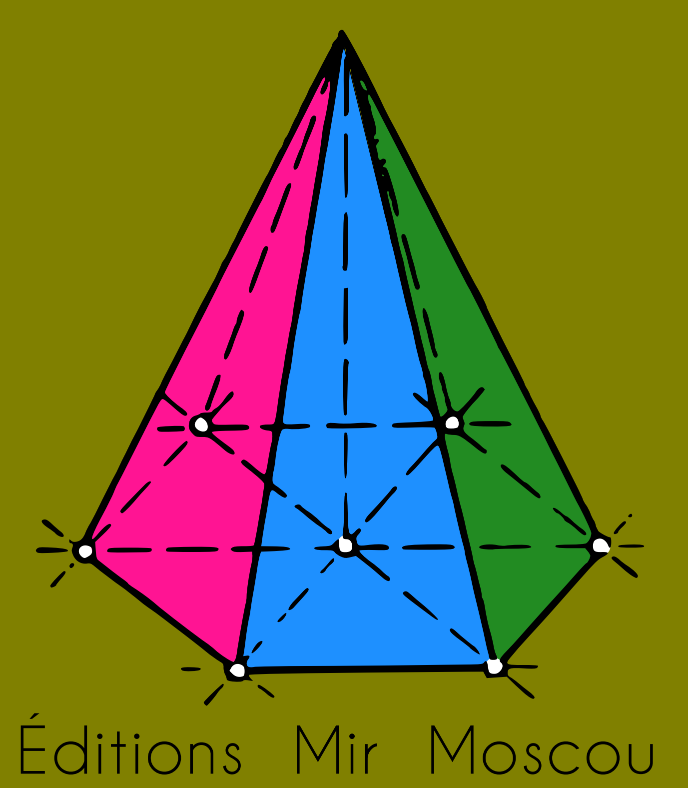
G. Marchouk, V. Shaydourov

RAFFINEMENT DES SOLUTIONS DES SCHÉMAS AUX DIFFÉRENCES



Éditions Mir Moscou

Г. И. МАРЧУК, В. В. ШАЙДУРОВ

ПОВЫШЕНИЕ ТОЧНОСТИ РЕШЕНИЙ РАЗНОСТНЫХ СХЕМ

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА

G. MARCHOUK, V. SHAYDOUROV

RAFFINEMENT DES SOLUTIONS DES SCHÉMAS AUX DIFFÉRENCES

ÉDITIONS MIR ● MOSCOU

Traduit du russe
par Irina PÉTROVA

На французском языке

PRÉFACE À L'ÉDITION RUSSE

Cette monographie est écrite à la suite de la nécessité toujours plus pressante de résoudre les problèmes avec une précision élevée. Malgré un essor impétueux du matériel de calcul électronique, on ne dispose malheureusement pas de moyens indispensables pour y arriver par les méthodes numériques simples peu exactes.

La rapidité des ordinateurs augmente sans cesse, mais la capacité des mémoires limite la dimension des tableaux à traiter, ce qui empêche une mise en œuvre efficace des algorithmes. Or, les opérations de calcul et les opérations logiques réalisées avec une bonne efficacité jointes aux mémoires suffisamment grandes permettent de diminuer de plusieurs (sic !) ordres le temps de résolution des problèmes réduits à des blocs standards.

Cela remet au premier plan les méthodes économiques au possible donnant lieu à des tableaux peu importants qui ne dépassent pas les possibilités des processus spécialisés.

Les progrès réalisés dans les domaines de l'aérodynamique, de la dynamique des fluides, de la théorie de l'échange de masse, de la théorie du transfert de rayonnement et dans d'autres branches scientifiques et techniques ont conduit à poser de nouvelles classes de problèmes et à élaborer des modèles mathématiques inédits qui décrivent toute une gamme d'effets physiques fins. La résolution des problèmes se ressent de ces effets qui influencent donc, dans de nombreux cas, la conception des machines à calculer.

L'étude de la stabilité des écoulements fluides autour des corps relève, par exemple, des équations avec viscosité petite. D'autre part, la mise en œuvre numérique introduit dans le modèle la « viscosité artificielle » qui est en général beaucoup plus grande que la viscosité physique et qui altère sérieusement l'aspect physique des phénomènes.

Il en est de même des problèmes relatifs à la dynamique de l'Océan mondial en régime faiblement turbulent. La liste des exemples pourrait être prolongée.

Un moyen de résolution naturel est en l'occurrence un algorithme de calcul qui garantit une grande précision des approximations.

Les méthodes de raffinement des solutions occupent une place honorable dans la littérature mathématique. C'est le procédé simple consistant à diminuer de façon proportionnelle les intervalles de discrétisation des problèmes différentiels, les schémas aux différences à plusieurs points et l'extrapolation de Richardson basée sur plusieurs solutions associées à une suite de réseaux. C'est cette dernière technique qui fournit la matière de notre monographie car elle nous paraît le moyen de raffinement le plus puissant et universel.

Certaines propriétés de la méthode de Richardson en font un des principaux instruments de l'Analyse numérique. C'est, *primo*, le fait d'utiliser les approximations aux différences simples des problèmes différentiels, *secundo*, un fonctionnement uniforme des algorithmes pour les réseaux avec paramètres différents, et, *tertio*, la logique simple de l'algorithme dans son ensemble.

La méthode est donc devenue l'objet du plus vif intérêt, mais de nombreux ouvrages traitant de l'extrapolation de Richardson (et des méthodes de Runge et de Romberg qui en sont des modifications) ne contiennent pas un seul exposé systématique de celle-ci et de ses applications éventuelles à diverses classes de problèmes.

Notre monographie est censée combler partiellement cette lacune.

Le lecteur y trouvera, en plus des aspects généraux de l'extrapolation de Richardson pour des problèmes linéaires abstraits, quelques-unes de ses variantes concrètes. S'agissant des équations différentielles ordinaires linéaires, la justification de la méthode dérive de règle des résultats abstraits. Quant aux équations aux dérivées partielles, elles soulèvent plusieurs problèmes auxiliaires. Dans le cas elliptique, c'est avant tout la régularité des solutions, tout spécialement leur comportement asymptotique au voisinage des points anguleux. Dans le cas parabolique, il y a imbrication entre la régularité et les procédés de décomposition du problème à plusieurs variables en une suite de problèmes en dimension un. On donne les résultats de l'extrapolation de Richardson pour le problème de valeurs propres et des équations quasi linéaires.

Les auteurs n'ont pas négligé non plus les systèmes d'équations algébriques à matrices dégénérées, et des problèmes avec couche limite leur ont permis de développer la méthode de l'extrapolation sur un paramètre petit.

Construire les solutions approchées avec une précision élevée constitue un problème actuel de l'Analyse numérique. La présente monographie est une synthèse des travaux effectués par les auteurs au cours de la dernière décennie. Ces travaux concernent essentiellement l'extrapolation de Richardson en tant que procédé de raffinement des solutions numériques des problèmes qui se posent en physique mathématique. La méthode a fait ses preuves dans beaucoup d'applications, et elle sert d'outil puissant dans des cas compliqués. Quelques-unes de ses modifications simples ont été incluses dans les cours d'Analyse numérique donnés aux étudiants des Universités de Novossibirsk et de Krasnoïarsk (voir G. Marchouk, *Méthodes de calcul numérique*, Moscou, 1977*).

On n'a regroupé dans la Bibliographie *in fine* que les livres et les articles dont les résultats ont été utilisés par les auteurs. Nombre d'ouvrages intéressants se rapportant aux questions traitées ne figurent donc pas sur la liste.

Lorsque nous avons étudié le problème annoncé par le titre de cette monographie, nous avons bénéficié des hautes compétences de nombreux analystes numériciens. Le livre a pris sa forme grâce aux discussions que nous avons eues avec E. Volkov, Y. Kouznétsov, V. Lébédev, J.-L. Lions, V. Agochkov, A. Matsokine. Qu'ils veuillent bien trouver ici le témoignage de notre gratitude.

Nous exprimons nos remerciements à G. Démidov, Y. Valitski, V. Sapojnikov et V. Shtchépanovski qui ont lu le manuscrit et fait plusieurs suggestions de valeur. Nous sommes reconnaissants à A. Alexéev, T. Babourina, B. Bagaev, S. Bersénev, B. Dobrontz et A. Joukov d'avoir pris part à la réduction technique du manuscrit et effectué les essais numériques.

G. Marchouk, V. Shaydourov

* La traduction française a paru aux Editions de Moscou en 1980. (*N.d.R.*)

PRÉFACE À L'ÉDITION FRANÇAISE

Depuis que ce livre a paru en langue russe, les auteurs ont réuni plusieurs nouveaux exemples d'emploi des algorithmes. On obtenait un effet particulièrement spectaculaire dans les problèmes pluridimensionnels et quasi linéaires de la physique mathématique lorsque les schémas aux différences (ou variationnels aux différences) simples doublés de l'extrapolation de Richardson ou le raffinement par des différences d'ordre relativement peu élevé fournissaient des résultats très concluants en ce qui concerne la précision. Car nombreux sont les cas où seuls les schémas simples d'ordre d'approximation faible possèdent des propriétés requises telles que la stabilité asymptotique, la monotonie, la validité des analogues discrets des lois de conservation, l'économie d'un pas de temps, la symétrie, l'algorithme explicite, etc. Ces schémas garantissent souvent un comportement qualitatif juste de la solution approchée sans qu'on obtienne une bonne précision même si l'on travaille avec de grandes capacités. Mais s'il y a régularité ou si l'on supprime spécialement les singularités, alors l'extrapolation de Richardson ou la correction des solutions des schémas par des différences d'ordre assez peu élevé aboutissent nécessairement à des résultats très précis.

L'ouvrage que nous présentons aux lecteurs de langue française diffère légèrement de l'édition russe: les auteurs ont décrit sous forme générale la méthode des différences d'ordre supérieur pour les problèmes linéaires, ajouté des numéros traitant des procédés d'extrapolation non linéaires, de l'influence des erreurs de calcul et remplacé certains algorithmes d'extrapolation par d'autres plus universels.

La manière d'exposer les matériaux reste la même. S'agissant des problèmes linéaires, tous les résultats découlent de règle des théorèmes généraux, et on démontre chaque cas non linéaire sans négliger aucun calcul intermédiaire.

G. Marchouk, V. Shaydourov

INTRODUCTION

Avec une large utilisation des ordinateurs dans les problèmes de la science et de la technique, de nombreux algorithmes sont apparus dont la plupart consistent à ramener la problème différentiel primitif à des problèmes d'algèbre linéaire. Il n'existe pas pour le moment de procédé plus universel pour traiter les problèmes soulevés par les applications. La dimension du problème d'algèbre linéaire dépend naturellement du paramètre de réduction qui est d'ordinaire le pas du réseau de discrétisation. Cela étant, on garantit une précision d'autant plus grande que le paramètre est petit. Lorsqu'on diminue celui-ci, le nombre d'équations linéaires augmente en conséquence et le volume de calcul monte en flèche.

Malgré la capacité toujours plus grande des calculateurs, les algorithmes ne sont jamais aussi exacts qu'on le désire. On cherche donc des moyens universels économiques pour créer les algorithmes de calcul et les exécuter sur machine.

Il est connu que la précision des solutions numériques est de règle proportionnelle à une puissance du paramètre de discrétisation et que dans le cas des schémas le plus simples et économiques, elle l'est à deux premières puissances.

Au début du XXe siècle, Richardson a proposé un procédé foncièrement nouveau pour améliorer la précision des solutions numériques des problèmes linéaires. Il a décrit des schémas d'algorithmes où l'on fait le calcul pour plusieurs pas différents. Il se trouve que sous certaines conditions la combinaison linéaire des solutions ainsi obtenues est d'ordre de précision plus élevé. L'idée de combiner les solutions associées à divers paramètres deviendra la clef de voûte des algorithmes de calcul pour de nombreux problèmes relevant des équations différentielles.

Malheureusement, la méthode de Richardson (et les procédés similaires dus à Runge et à Romberg) a été longtemps utilisée de

manière plutôt heuristique sans aucune motivation sérieuse, si bien que son usage abusif dans certains cas a fait douter de son utilité pratique. Aujourd'hui, la méthode est remise sur le tapis. Et cela pour les raisons suivantes.

Lorsqu'on élabore une procédure de calcul économique, on utilise largement l'information à priori sur la solution du problème et la classe de fonctions dont celle-ci est élément. Soit, par exemple, un problème qui admet, on le sait, comme solution une parabole. Dans ce cas, il suffit, bour la trouver, de connaître ses valeurs en trois points. Et c'est avec l'hypothèse à priori qu'on la définit en tous les points restants. Cet exemple simple montre on ne peut mieux que l'information à priori permet de réduire notablement la dimension des tableaux sans qu'une perte d'information s'ensuive.

S'agissant d'une équation différentielle, l'information à priori décisive est le degré de régularité de la solution, son comportement asymptotique dans les situations extrémales, sa dépendance vis-à-vis des données du problème, etc. Cela étant, plus elle est complète, moins la classe fonctionnelle correspondante est riche, ce qui facilite la recherche de la solution.

Il se peut que l'information à priori soit si abondante que la classe en question se réduit à un seul élément qui est visiblement la solution cherchée. Mais c'est là le cas limite. Dans la pratique, on possède en général d'utiles renseignements sur la solution qui définissent de règle l'espace fonctionnel nécessaire. Il arrive que c'est l'information à priori qui détermine le choix de la méthode de résolution approchée.

Supposons, en effet, qu'on cherche formellement la solution d'une équation différentielle par une méthode numérique exacte à l'ordre 1 par rapport au pas h du réseau. Si l'on discrétise avec le paramètre h, puis avec $h/2$, alors on dit en général que la seconde solution est deux fois plus précise que la première. Si l'on prend $h/3$, le résultat est trois fois plus précis, et ainsi de suite. On en conclut qu'avec notre hypothèse à priori sur l'ordre de la méthode, l'exactitude du résultat est approximativement proportionnelle au pas du réseau. Si l'on veut une solution 10^2 fois plus précise que celle associée à h, on réduit donc le pas de 10^2 fois. On conçoit que la chose n'est pas toujours possible même si l'on travaille avec des capacités très grandes. C'est surtout vrai des problèmes à deux

et à trois variables car la dimension des tableaux est alors élevée au carré et au cube respectivement.

On fait appel à une information auxiliaire. On suppose que la solution du problème différentiel est plus régulière, ce qui entraîne, pour la solution approchée, l'existence de trois dérivées bornées par rapport à h. On utilise le développement taylorien par rapport à h et on démontre qu'avec la combinaison linéaire de trois solutions associées à h, $h/2$ et $h/3$, on aboutit à une précision sensiblement plus élevée. Dans le cas considéré, elle est $O(h^3)$. Si l'on prend, par exemple, le paramètre sans dimension $h=1/10$, la précision sera de l'ordre de 10^{-3}, et ce résultat sera atteint moyennant trois solutions approchées correspondant aux réseaux successifs de pas h, $h/2$, $h/3$. On note que ce procédé de raffinement nous évite de résoudre le problème approché pour le pas de discrétisation $\sim 10^{-3}$, ce qui conduit en dimension un aux nœuds en nombre proportionnel à 10^3. Or, le procédé décrit nous a permis de nous borner à trois solutions et aux nœuds dont le nombre est proportionnel à 10, 20 et 30 respectivement. En dimension deux et trois, la différence est encore plus spectaculaire.

Généralement parlant, il s'agit d'un analogue du problème de définir une parabole connaissant ses trois points quelconques à la différence que les « points » sont ici les solutions approchées de paramètres h, $h/2$, $h/3$ donnés et qu'au lieu d'obtenir la solution exacte du problème différentiel, on aboutit à une solution approchée d'ordre de précision $O(h^3)$. Avec l'hypothèse à priori de régularité plus grande, on élargit naturellement l'ensemble des solutions approchées, ce qui donne une précision plus élevée sur le résultat définitif. L'idée de base de l'extrapolation de Richardson est justement d'utiliser un jue de solutions approchées associées à plusieurs réseaux successifs.

La présente monographie se propose entre autres de décrire des façons plus ou moins générales dont on met en œuvre la méthode de Richardson pour une vaste classe de problèmes de la physique mathématique et de donner les conditions de sa validité. S'agissant des problèmes linéaires, ces conditions s'énoncent sous forme générale, si bien que la tâche du mathématicien se réduit dans chaque cas concret à établir de façon constructive si elles sont oui ou non satisfaites. Ces conditions permettent de formuler les théorèmes de convergence généraux du *Chapitre premier*.

Quatre chapitres suivants comprennent autant de paragraphes que de types de problèmes différentiels. Et cela pour deux raisons suivantes. *Primo*, les méthodes numériques portent l'empreinte des propriétés du problème posé, et même le type du problème différentiel s'avère des fois décisif lorsqu'on choisit une méthode économique. *Secundo*, les résultats théoriques de la monographie utilisent essentiellement l'information à priori sur la régularité des données connues (et non sur celle de la solution même), ce qui a obligé les auteurs à emprunter à la théorie des équations différentielles d'utiles renseignements complémentaires relatifs à la compatibilité de ces propriétés des données et de la solution. Cette information intervient naturellement au fur et à mesure des besoins.

Dans la méthode de Richardson, on obtient en principe des solutions améliorées de tout ordre de précision si l'on est dans les conditions de concordance et de régularité correspondantes. Ces questions sont traitées dans les *Chapitres* 3, 4 et 5 aussi bien pour des problèmes relativement simples que pour des cas compliqués tels que les équations à coefficients discontinus et les problèmes dont les solutions possèdent des singularités en certains points frontières. Il est connu que la construction des solutions approchées de haute précision s'avère particulièrement ardue quand la frontière du domaine présente des points anguleux. Ce cas est étudié dans la *Chapitre* 4 consacré aux équations elliptiques. La solution a d'ordinaire des singularités au voisinage des points anguleux, et on y commet une erreur d'approximation qui n'est pas compatible avec l'erreur faite loin de ces points. Formellement, la méthode de Richardson est en défaut. Mais si l'on dégage au préalable les singularités sous forme explicite et si l'on représente la solution comme combinaison des solutions singulières et de la partie régulière restante, il y a intérêt à extrapoler à la Richardson pour cette dernière, et le problème est susceptible d'être résolu avec une précision élevée.

La pratique de la méthode de Richardson dans le cas linéaire a suggéré aux auteurs de raffiner de même les solutions des problèmes non linéaires. Il est vrai que les conditions à priori imposées à l'opérateur, aux entrées et à la solution deviennent plus sévères, mais les applications éventuelles de la méthode sont très nombreuses.

Le fait que la méthode de Richardson fonctionne pour des problèmes non stationnaires de la physique mathématique, tout spécialement pour des équations du type parabolique, nous paraît fort

important. Ces dernières années, on a élaboré pour la résolution approchée de ces problèmes une quantité de bons algorithmes qui sont décrits, par exemple, dans une monographie de l'un des auteurs. On signale l'intérêt capital de la méthode de décomposition qui permet de représenter un gros problème à plusieurs variables par une suite de problèmes en dimension un qui sont calculés efficacement sur ordinateur. On a pensé d'abord que les erreurs d'approximation dans la décomposition entravent, voire interdisent, la méthode de Richardson pour approcher les problèmes avec une grande précision. Or, on montre dans la *Chapitre* 5 qu'il n'en est rien. Il y a plus. La méthode de décomposition procède de façon classique sans qu'on ait besoin de modifier l'algorithme. Si les auteurs n'ont réussi à légitimer l'extrapolation de Richardson que dans le cas d'un seul schéma, ce dernier présente par contre l'avantage d'être très employé en calcul automatique. D'ailleurs, on a des raisons de penser que la classe de ces schémas deviendra de plus en plus fournie.

L'idée de Richardson n'a été pas utile que dans les problèmes différentiels. Elle inspire, par exemple, plusieurs algorithmes pour des équations algébriques linéaires dégénérées (*Ch.* 6), qui continuent en quelque sorte la méthode de régularisation de Tikhonov. On remplace le système algébrique par un système voisin non dégénéré avec un paramètre ε qu'on fait tendre à la limite vers 0. Dans un premier temps, on choisit ε tel qu'on obtienne un système d'équations bien conditionné et une solution qui n'est pas trop grossièrement approchée. On aborde ensuite les cas avec $\varepsilon/2$, $\varepsilon/3$, ... Il se trouve que la combinaison linéaire des solutions munies de certains poids donne la solution normale dont la précision est $O(\varepsilon^n)$, avec n le nombre de problèmes auxiliaires associés à ε, $\varepsilon/2$, ..., ε/n. On traite de sorte de nombreux problèmes d'algèbre linéaire (en particulier, les problèmes qui relèvent de la méthode de pénalisation), les problèmes différentiels dont les dérivées d'ordre supérieur sont affectées d'un paramètre petit, et ainsi de suite. Ces problèmes interviennent, par exemple, en physique de l'atmosphère et de l'océan en régime faiblement turbulent, dans le cas d'écoulement des fluides avec les nombres de Reynolds importants, etc.

On a regroupé dans le *Chapitre* 7 plusieurs résultats auxiliaires qui étayent les raisonnements théoriques des chapitres précédents et que les spécialistes de l'Analyse numérique connaissent très bien.

Dans tous les essais numériques, les erreurs d'arrondi et les erreurs commises en calculant les fonctions élémentaires en dépassent pas le dernier chiffre significatif du résultat. Cela permet d'apprécier le gain en précision indépendamment du type de la machine, de celui du traducteur du langage algorithmique en langage machine et de l'art du programmeur. L'exemple cité à la fin du § 2.1 est une exception à la règle (on y étudie précisément l'influence des erreurs d'arrondi sur les résultats de l'extrapolation).

Les exemples numériques du livre mettent en relief la caractère universel de la méthode de Richardson. En effet, on atteint un résultat très précis en résolvant une succession de problèmes aux différences avec des schémas d'approximation de même structure. Seuls changent le paramètre de réduction et le pas du réseau, et le nombre d'équations linéaires approximant le problème augmente naturellement. L'invariance de structure des algorithmes est une condition *sine qua non* de la création des softwares qui permettent d'élaborer des paquets de programmes généraux.

CHAPITRE PREMIER

GÉNÉRALITÉS

Ce chapitre débute par un exemple dans lequel les solutions discrètes d'un schèma d'ordre d'approximation peu élevé permettent d'augmenter la précision de la solution approchée par l'extrapolation de Richardson ou les différences d'ordre supérieur. Les résultats sont généralisés à des problèmes abstraits: on énonce des conditions suffisantes pour que la solution approchée admette un développement suivant le pas de discrétisation et on démontre qu'avec ce développement, on utilise avec succès l'extrapolation de Richardson et le raffinement par les différences d'ordre supérieur.

1.1. Un exemple simple

Dans les pages qui suivent, nous allons résoudre une équation différentielle linéaire du premier ordre. Ce sera un exemple modèle dans lequel on extrapolera par Richardson au moyen d'un ensemble de solutions aux différences peu précises, et on justifiera sommairement l'ordre de précision plus élevé de la solution extrapolée. On donnera ensuite une variante simple de la méthode des différences d'ordre supérieur et les résultats numériques caractéristiques de l'efficacité pratique des algorithmes.

1.1.1. Extrapolation de Richardson locale et global

Soit l'équation différentielle

$$u' + u = f \quad \text{sur} \quad (0,\ 1) \tag{1.1}$$

avec la condition initiale

$$u(0) = u_0. \tag{1.2}$$

On suppose que f est indéfiniment dérivable sur $[0,\ 1]$.

Pour que le problème reçoive une solution numérique, on construit le réseau régulier

$$\bar{\omega}_\tau = \{t_j = j\,\tau; \qquad j = 0,\ 1,\ \ldots,\ M\} \tag{1.3}$$

de pas $\tau = 1/M$ ($M \geqslant 2$ étant un entier) et on introduit les points médians

$$\bar{\omega}_\tau = \{t_{j+1/2} = (j + 1/2)\,\tau; \qquad j = 0,\ 1,\ \ldots,\ M - 1\}. \tag{1.4}$$

On remplace suivant le schéma de Crank-Nicholson l'équation (1.1) associée aux points médians par le système d'équations algébriques

$$u_{\bar{t}}^{\tau} + u_{\tilde{t}}^{\tau} = f \quad \text{sur} \quad \bar{\omega}_\tau. \tag{1.5}$$

Ici

$$u_{\bar{t}}(t) = \frac{u(t + \tau/2) - u(t - \tau/2)}{\tau}, \quad u_{\tilde{t}}(t) = \frac{u(t + \tau/2) + (u\,t - \tau/2)}{2}.$$

On adjoint à (1.5) la condition

$$u^\tau(0) = u_0, \tag{1.6}$$

conséquence de (1.2). Le problème proposé admet pour solution la fonction discrète u^τ qui approche u avec une précision d'ordre 2 (voir [65]) en tous les nœuds de $\bar{\omega}_\tau$:

$$\| u^\tau - u \|_{C,\tau} \leqslant c_1 \tau^2 \,* \tag{1.7}$$

(voir Notations *in fine*).

On définit l'erreur comme différence de la solution approchée (discrète) u^τ et de la solution exacte u aux nœuds $t \in \bar{\omega}_\tau$, et on montre qu'on a pour $\tau \to 0$

$$u^\tau - u = \tau^2 v + \eta^\tau \quad \text{sur} \quad \bar{\omega}_\tau, \tag{1.8}$$

où v est une fonction régulière définie sur le segment $[0,\ 1]$ et indépendante de τ et η^τ une fonction discrète définie sur $\bar{\omega}_\tau$ qui prend des valeurs $O(\tau^4)$. La notation $O(\tau^4)$ a un sens usuel: étant donnée une fonction discrète φ quelconque, l'égalité $\varphi = O(\tau^k)$ sur l'ensemble D équivaut à l'existence d'une constante $c_0 \in [0, \infty]$ telle qu'on ait $|\varphi| \leqslant c_0 \tau^k$ sur D.

On suppose que le développement (1.8) a lieu, auquel cas

$$u^\tau = u + \tau^2 v + \eta^\tau \quad \text{sur} \quad \bar{\omega}_\tau.$$

Ces égalités aidant, on récrit (1.5) et (1.6):

$$u_{\bar{t}} + \tau^2 v_{\bar{t}} + \eta_{\bar{t}}^\tau + u_{\tilde{t}} + \tau^2 v_{\tilde{t}} + \eta_{\tilde{t}}^\tau = f \quad \text{sur} \quad \bar{\omega}_\tau, \tag{1.9}$$

$$u(0) + \tau^2 v(0) + \eta^\tau(0) = u_0. \tag{1.10}$$

On développe u et v de (1.9) en formule de Taylor par rapport au point t, il vient

$$u' + u + \tau^2 \left(\frac{1}{24} u''' + \frac{1}{8} u'' + v' + v\right) + O(\tau^4) + \eta_{\bar{t}}^\tau + \eta_{\tilde{t}}^\tau = f.$$

---

* Dans ce livre, c_i (i entier) désigne des constantes indépendantes de τ, t, x, h.

Comme les coefficients de τ^0 et τ^2 ne dépendent pas de τ et les autres termes du premier membre sont probablement d'ordre infinitésimal supérieur, on a en tous les points, par suite de l'arbitraire laissé sur τ,

$$u' + u = f, \tag{1.11}$$

$$\frac{1}{24} u''' + \frac{1}{8} u'' + v' + v = 0. \tag{1.12}$$

On a de meme pour les données initiales (1.10):

$$u(0) = u_0, \tag{1.13}$$

$$v(0) = 0. \tag{1.14}$$

Les deux premières relations s'interprètent naturellement comme étant des équations en u et v, et les deux dernières comme des conditions initiales. La fonction u étant solution du problème (1.1), (1.2) les vérifie nécessairement. Quant à l'information de valeur obtenue au sujet de v, elle nous sera utile dans la suite.

On note que v est solution de l'équation

$$v' + v = -\frac{1}{24} u''' - \frac{1}{8} u'', \qquad t \in (0, 1), \tag{1.15}$$

avec la condition initiale

$$v(0) = 0. \tag{1.16}$$

Il est évident que ce problème est possible et admet une solution unique

$$v(t) = -\frac{1}{24} \int_0^t e^{x-t} (u'''(x) + 3 u''(x))\, dx \tag{1.17}$$

qui possède des dérivées de tous ordres sur [0, 1]. Il y a lieu de dire que la fonction v ainsi construite ne dépend pas du paramètre τ.

On montre la propriété de borne uniforme de η^τ discrète définie aux nœuds de $\bar{\omega}_\tau$ par la relation

$$\eta^\tau(t) = u^\tau(t) - u(t) - \tau^2 v(t). \tag{1.18}$$

Soit

$$\eta_{\bar{t}}^{\tau} + \eta_{t}^{\tau} = \left(\frac{1}{\tau} + \frac{1}{2}\right) \eta^\tau\left(t + \frac{\tau}{2}\right) - \left(\frac{1}{\tau} - \frac{1}{2}\right) \eta^\tau\left(t - \frac{\tau}{2}\right).$$

d'où l'on tire compte tenu de (1.18):

$$\eta_{\bar{t}}^{\tau} + \eta_{t}^{\tau} = u_{\bar{t}}^{\tau} + u_{t}^{\tau} - (u_{\bar{t}} + u_{t}) - \tau^2 (v_{\bar{t}} + v_{t}), \quad t \in \omega_\tau.$$

On fixe un nœud $t \in \omega_\tau$ et on transforme le second membre. En vertu de (1.5), la somme de deux premiers termes vaut $f(t)$. On

rapporte les fonctions des termes suivants aux valeurs en t à l'aide du développement taylorien avec reste sous forme de Lagrange (lemme 1.1, § 7.1), il vient

$$\eta^{\tau}_{\bar t} + \eta^{\tau}_{\bar t} = (f - u' - u) - \tau^2 \left(\frac{1}{24} u''' + \frac{1}{8} u'' + v' + v\right) + \xi,$$

où

$$|\xi| \leqslant c_2 \tau^4 \quad \forall\, t \in \bar\omega_\tau.$$

Les équations (1.1) et (1.15) conduisent à

$$\eta^{\tau}_{\bar t} + \eta^{\tau}_{\bar t} = \xi \quad \text{sur} \quad \bar\omega_\tau. \tag{1.19}$$

On en déduit la relation

$$|\eta^\tau (t + \tau/2)| \leqslant \eta^\tau |(t + \tau/2)| + \tau |\xi(t)|.$$

Il est immédiat de vérifier que $\eta^\tau(0) = 0$ en vertu des conditions (1.2), (1.6) et (1.16). En raisonnant par récurrence, on établit donc l'estimation

$$\|\eta^\tau\|_{C,\tau} \leqslant \sum_{\bar\omega_\tau} |\xi|_\tau \quad \text{ou} \quad \|\eta^\tau\|_{C,\tau} \leqslant c_2 \tau^4. \tag{1.20}$$

On note que (1.19) et (1.20) entraînent, pour la dérivée aux différences,

$$\max_{\bar\omega_\tau} |\eta^{\tau}_{\bar t}| \leqslant \max_{\bar\omega_\tau} |\eta^{\tau}_{\bar t}| + O(\tau^4) \leqslant O(\tau^4). \tag{1.21}$$

On va décrire deux méthodes de raffinement basées sur le développement démontré. On construit les réseaux (1.4) de pas τ et $\tau/2$ et on cherche pour chaque réseau la solution du problème approché (1.5), (1.6). Soit u^τ et $u^{\tau/2}$ deux fonctions discrètes qui sont les solutions demandées (la précision de chacune étant en τ^2). On forme une combinaison linéaire des solutions approchées relatives aux nœuds du réseau de pas τ:

$$U(t) = \frac{4}{3} u^{\tau/2}(t) - \frac{1}{3} u^\tau(t), \qquad t \in \bar\omega_\tau, \tag{1.22}$$

et on montre que U approche u avec une précision de l'ordre de τ^4.

On a en chaque point de $\bar\omega_\tau$

$$u^\tau(t) = u(t) + \tau^2 v(t) + O(\tau^4),$$
$$u^{\tau/2}(t) = u(t) + (\tau^2/4)\, v(t) + O(\tau^4),$$

d'où

$$U(t) = u(t) + O(\tau^4). \tag{1.23}$$

Ainsi, la solution améliorée (1.22), combinaison linéaire des solutions approchées (exactes à l'ordre 2 en τ) du problème (1.5), (1.6), approche la solution exacte u aux nœuds de $\bar{\omega}_\tau$ à l'ordre 4 en τ.

Ce procédé de raffinement qui consiste à construire une solution approchée à partir des fonctions discrètes, solutions moins précises des problèmes approchés, a été proposé par Richardson (voir [122], 123]) qui lui a donné le nom *d'extrapolation à la limite* (deferred, approach to the limit), la valeur limite étant

$$u(t) = \lim_{\tau \to 0} u^\tau(t).$$

On l'appellera plus tard *l'extrapolation de Richardson*. Il s'agit d'une méthode *globale* parce qu'on extrapole pour plusieurs solutions approchées connues. Il existe par contre un procédé d'extrapolation qu'on pourrait appeler *local*. En effet, il consiste à extrapoler à chaque pas, et on initialise chaque fois avec l'approximation précédente. Soit, par exemple, $w^\tau(x)$ la solution associée au point x. On pose

$$v^\tau(x) = w^\tau(x), \qquad v^{\tau/2}(x) = w^\tau(x). \tag{1.24}$$

On opère selon le schéma de Crank-Nicholson, et on effectue un pas de longueur τ pour v^τ et deux pas de longueur $\tau/2$ pour $v^{\tau/2}$. On a

$$v^\tau_{\bar{t}}(t) + v^\tau_{\hat{t}}(t) = f(t), \qquad t = x + \tau/2;$$

$$v^\tau_{\bar{t}}(t) + v^{\tau/2}_{\hat{t}}(t) = f(t), \qquad t = x + \tau/4, x + 3\tau/4.$$

On fait la somme et on utilise les mêmes poids que dans le cas global:

$$w^\tau(x + \tau) = 4/3\, v^{\tau/2}(x + \tau) - 1/3\, v^\tau(x + \tau). \tag{1.25}$$

Avec l'approximation initiale $w^\tau(0) = u_0$, on trouve en principe par (1.24) (1.25) la solution approchée w^τ aux nœuds du réseau $\bar{\omega}_\tau$.

En dépit de leur ressemblance apparente, ces deux algorithmes présentent des différences profondes sur le plan théorique. Le premier conserve la stabilité des schémas aux différences de départ, tandis qu'on utilise en fait dans le second le principe de Runge présidant à la construction des méthodes d'approximation d'ordre élevé, si bien qu'il conduit à un schéma du type Runge-Kutta qualitativement différent. La stabilité de cette méthode n'est plus une conséquence nécessaire de celle du schéma de départ. L'exemple ci-dessus illustre bien nos paroles. Dans cet exemple, le problème 1.1) est abordé par le schéma de Crank-Nicholson stable quel que

soit $r > 0$. D'autre part, l'algorithme (1.24), (1.25) (voir [79]) est instable pour τ supérieurs à un certain τ_0. On trouve ce résultat dans [130], ainsi qu'une modification de l'extrapolation de Richardson locale qui ne présente pas cet inconvénient.

On confond des fois deux mises en œuvre décrites, ce qui explique le caractère contradictoire de certaines recommandations pratiques.

Dans la suite, on s'occupera du cas global seul et on dira «extrapolation» tout court.

1.1.2. Méthode des différences d'ordre supérieur

Il existe une autre méthode pour améliorer la précision de la solution discrète du problème (1.5), (1.6), qui est également basée sur le développement (1.8). Elle est connue sous le nom de ***méthode des différences d'ordre supérieur*** ou de ***méthode des approximations successives*** (the difference (or deferred) correction method).

On modifie au préalable l'équation aux différences (1.5) de façon d'augmenter son ordre d'approximation:

$$L^\tau v^\tau \equiv v_{\bar{t}}^\tau + v_{\bar{t}}^\tau - \tau^2 \left(\frac{1}{24} v_{\bar{t}\bar{t}\bar{t}}^\tau + \frac{1}{8} v_{\bar{t}\bar{t}\bar{t}}^\tau \right) = f$$

$$\forall\, t \in \bar{\omega}_\tau \setminus \{\tau/2,\ 1 - \tau/2\}. \tag{1.26}$$

On a, avec les notation fondamentales,

$$v_{\bar{t}\bar{t}}(t) = (v_{\bar{t}}(t))_{\bar{t}} = \frac{1}{\tau} (v_{\bar{t}}(t + \tau/2) - v_{\bar{t}}(t - \tau/2)) =$$

$$= \frac{1}{\tau^2} (v(t + \tau) - 2v(t) + v(t - \tau)),$$

$$v_{\bar{t}\bar{t}\bar{t}}(t) = \frac{1}{2} (v_{\bar{t}\bar{t}}(t + \tau/2) + v_{\bar{t}\bar{t}}(t - \tau/2)) =$$

$$= \frac{1}{2\pi^2} (v(t + 3\tau/2) - v(t + \tau/2) - v(t - \tau/2) + v(t - 3\tau/2)),$$

et ainsi de suite.

On vérifie sans peine à l'aide du développement taylorien que l'opérateur aux différences L^τ du premier membre de (1.26) approche l'opérateur différentiel de (1.1) avec une précision du quatrième ordre:

$$L^\tau u = u' + u + O(\tau^4) \qquad \forall\, t \in \bar{\omega}_\tau \setminus \{\tau/2,\ 1 - \tau/2\}. \tag{1.27}$$

On note que l'équation à 4 points (1.27) n'est pas juste pour les nœuds $\tau/2$ et $1 - \tau/2$ puisqu'elle contient dans ce cas les valeurs

de v^τ aux points $-\tau$, $1+\tau$ extérieurs au segment $[0,1]$. S'agissant de ces nœuds, on écrit donc les équations aux différences

$$L^\tau v^\tau(\tau/2) \equiv v_t^\tau(\tau/2) + v_{\bar t}^\tau(\tau/2) -$$
$$- \frac{\tau^2}{24} v_{\bar t t \bar t}(\tau/2) - \frac{\tau^2}{8} v_{\bar t t \bar t}(\tau/2) + \frac{\tau^3(3\tau-2)}{48} v_{\bar t t \bar t \bar t}^\tau(\tau) = f(\tau/2), \quad (1.28)$$

$$L^\tau v^\tau(1-\tau/2) \equiv v_t^\tau(1-\tau/2) + v_{\bar t}^\tau(1-\tau/2) -$$
$$- \frac{\tau^2}{24} v_{\bar t t \bar t}^\tau(1-\tau/2) - \frac{\tau^2}{8} v_{\bar t t \bar t}^\tau(1-\tau/2) +$$
$$+ \frac{\tau^3(3\tau+2)}{48} v_{\bar t t \bar t \bar t}^\tau(1-\tau) = f(1-\tau/2). \quad (1.29)$$

On vérifie de suite que les coefficients de $v^\tau(-\tau)$ et de $v^\tau(1+\tau)$ sont nuls et que l'erreur d'approximation est une quantité $O(\tau^3)$:

$$L^\tau u(t) = u'(t) + u(t) + O(\tau^3), \qquad t = \tau/2,\ 1-\tau/2. \quad (1.30)$$

On adjoint aux équations (1.26), (1.28), 1.29) une condition initiale découlant de (1.2), il vient le système d'équations algébriques linéaires

$$\begin{aligned} v^\tau(0) &= u_0, \\ L^\tau v^\tau &= f \qquad \forall\, t \in \bar\omega_\tau . \end{aligned} \quad (1.31)$$

De quelque façon qu'on numérote les inconnues et les équations, la matrice du système ne devient jamais triangulaire, ce qui entrave sérieusement la recherche de la solution. Il y a plus. La stabilité du système demande une démonstration spéciale.

On préfère donc à (1.31) une modification dépourvue de ces deux défauts. On résout le problème (1.5), (1.6). Connaissant u^τ, on cherche w^τ, solution de

$$w^\tau(0) = u_0, \quad (1.32)$$

$$w_t^\tau + w_{\bar t}^\tau = u_t^\tau + u_{\bar t}^\tau - L^\tau u^\tau + f \qquad \forall\, t \in \bar\omega_\tau. \quad (1.33)$$

Les deux problèmes ne diffèrent que par leurs seconds membres, et on aborde l'un et l'autre par un algorithme stable simple. On montre la validité de l'estimation

$$\| w^\tau - u \|_{C,\tau} \leqslant c_3 \tau^4. \quad (1.34)$$

En effet, u satisfait à la relation (1.8). On porte ce développement dans le second membre de (1.33), il vient pour $t \in \omega_\tau \setminus \{\tau/2, 1-\tau/2\}$

$$w_t^\tau + w_{\bar t}^\tau = \frac{\tau^2}{24} u_{\bar t t \bar t}^\tau + \frac{\tau^2}{8} u_{\bar t t \bar t}^\tau = \frac{\tau^2}{24} u_{\bar t t \bar t} +$$
$$+ \frac{\tau^2}{8} u_{\bar t t \bar t} + \frac{\tau^4}{24} v_{\bar t t \bar t} + \frac{\tau^4}{8} v_{\bar t t \bar t} + \frac{\tau^2}{24} \eta_{\bar t t \bar t}^\tau + \frac{\tau^2}{8} \eta_{\bar t t \bar t}^\tau + f. \quad (1.35)$$

On développe en série de Taylor pour évaluer les termes à droite:

$$u_{\bar t t t} = u''' + O(\tau^2), \qquad u_{\bar t t \bar t} = u'' + O(\tau^2), \tag{1.36}$$

$$|v_{\bar t t t}| \leqslant \frac{5}{4} \max_{[0,\,1]} |v'''| = O(1), \qquad |v_{\bar t t \bar t}| \leqslant \frac{5}{4} \max_{[0,\,1]} |v''| = O(1).$$

A la lumière des estimations (1.20), (1.21), on obtient

$$\begin{aligned} |\eta^{\tau}_{\bar t t t}| &\leqslant \frac{4}{\tau^2} \max_{\bar\omega_\tau} |\eta^{\tau}_{\bar t}| \leqslant O(\tau^2), \\ |\eta^{\tau}_{\bar t t \bar t}| &\leqslant \frac{4}{\tau^2} \max_{\bar\omega_\tau} |\eta^{\tau}_{\bar t}| \leqslant O(\tau^2). \end{aligned} \tag{1.37}$$

On transforme (1.35) compte tenu de (1.36), (1.37):

$$w^{\tau}_{\bar t} + w^{\tau}_{\bar t} = \frac{\tau^2}{24} u''' + \frac{\tau^2}{8} u'' + f + \zeta_1, \tag{1.38}$$

où

$$|\zeta_1(t)| \leqslant c_4 \tau^4 \qquad \forall\, t \in \bar\omega_\tau \setminus \{\tau/2,\ 1 - \tau/2\}. \tag{1.39}$$

On démontre un analogue de (1.38) pour $t = \tau/2$ et $t = 1-\tau/2$. Dans ce cas, le reste est évalué par

$$|\zeta_1(t)| \leqslant c_5 \tau^3, \qquad t = \tau/2,\ 1 - \tau/2. \tag{1.40}$$

Au numéro précédent, nous avons utilisé pour la solution exacte u le développement

$$u_{\bar t} + u_{\bar t} = u' + u + \frac{\tau^2}{24} u''' + \frac{\tau^2}{8} u'' + \zeta_2,$$

où

$$|\zeta_2(t)| \leqslant c_6 \tau^4 \qquad \forall\, t \in \bar\omega_\tau\,. \tag{1.41}$$

On le récrit moyennant l'équation (1.1), il vient

$$u_{\bar t} + u_{\bar t} = f + \frac{\tau^2}{24} u''' + \frac{\tau^2}{8} u'' + \zeta_2.$$

On retranche cette relation de (1.38) et on introduit la notation $\varepsilon^\tau = w^\tau - u$. On a

$$\varepsilon^{\tau}_{\bar t} + \varepsilon^{\tau}_{\bar t} = \zeta_1 - \zeta_2 \qquad \forall\, t \in \bar\omega_\tau\,. \tag{1.42}$$

Les conditions initiales (1.2), 1.6) entraînent

$$\varepsilon^\tau(0) = 0, \tag{1.43}$$

si bien que la solution du système (1.42), (1.43) admet un analogue de l'estimation (1.20) pour η^τ, à savoir

$$\|\zeta^\tau\|_{C,\,\tau} \leqslant \tau \sum_{\bar\omega_\tau} |\zeta_1 - \zeta_2|\,.$$

Avec les estimations (1.39) à (1.41), on est conduit à

$$\| \varepsilon^\tau \|_{C,\,\tau} \leqslant (c_1 + 2\,c_5 + c_6)\,\tau^4,$$

relation équivalente à (1.34).

Ainsi, on a démontré que la convergence de la solution du problème (1.32), (1.33) vers la solution exacte est d'ordre 4.

Cette tactique qui consiste à interpréter la méthode des différences d'ordre supérieur comme un procédé de mise en œuvre des schémas instables d'ordre d'approximation élevé (voir également [55]), nous paraît très constructive. Dans les paragraphes suivants du Chapitre premier, nous l'exposerons en termes généraux.

On note une différence d'avec l'extrapolation de Richardson: on construit cette fois les deux problèmes aux différences sur un même réseau. Cette propriété de la méthode ne dépend pas du problème différentiel et avantage celle-ci dans une certaine mesure lorsqu'il s'agit des problèmes pratiques. Par contre, le second membre de l'équation (1.33) est beaucoup plus compliqué que celui de (1.5), ce qui en détermine souvent un coût notablement plus grand en calcul automatique. •

1.1.3. Exemples numériques

Soit le problème

$$\begin{aligned} u' + u &= t\,(1+t)^{-2}, \qquad t \in (0,\,1), \\ u\,(0) &= 1. \end{aligned} \tag{1.44}$$

Il a pour solution la fonction $u(t) = (1+t)^{-1}$.

On a construit pour $M_k = 5 \cdot 2^{k-1}$, $k = 1, \ldots, 5$, les réseaux $\bar\omega_{\tau_k}$ et résolu sur chaque $\bar\omega_{\tau_k}$ le problème aux différences (1.5), (1.6). On a trouvé les erreurs maxima

$$\xi\,(M_k) = \| u^{\tau_k} - u \|_{C,\,\tau_k}.$$

En coordonnées logarithmiques, leur dépendance par rapport à M_k est représentée sur la fig. 1.1. On a formé pour chaque réseau de pas τ_k la combinaison linéaire de deux solutions

$$U^{\tau_k}\,(t) = \frac{4}{3}\, u^{\tau_k/2}(t) - \frac{1}{3}\, u^{\tau_k}\,(t)\,, \qquad t \in \bar\omega_{\tau_k},$$

et calculé l'erreur maximum

$$\varkappa_k = \| U^{\tau_k} - u \|_{C,\,\tau_k}. \tag{1.45}$$

On détermine naturellement l'efficacité de l'algorithme en rapportant (1.45) à $M_k + M_{k+1}$ car deux résolutions sur machine du problème approché (1.5), (1.6) nécessitent des opérations en nombre

proportionnel à $M_k + M_{k+1}$. En outre, $M_k + M_{k+1}$ caractérise le nombre de calculs du second membre, ce qui constitue un critère supplémentaire pour comparer les méthodes par différences.

On a mis en œuvre la méthode des différences d'ordre supérieur. On a trouvé par élimination de Gauss avec choix du pivot maximum les solutions v^{τ_k} du système « douteux » (1.31) pour cinq M_k cités et les erreurs maxima associées

$$\rho_k = \| v^{\tau_k} - u \|_{C,\ \tau_k}.$$

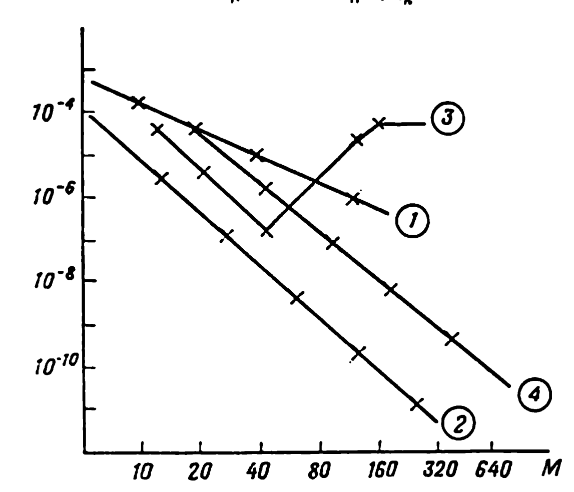


Fig. 1.1. Erreurs maxima sur les solutions approchées du problème (1.44)

1 — erreur sur la solution du schéma de Crank-Nicholson (1.5), (1.6); *2* - erreur sur la solution extrapolée (1.22); *3* — erreur sur la solution du problème «douteux» (1.31); *4* — erreur sur la solution du problème (1.32), (1.33) obtenue par la méthode des différences d'ordre supérieur

Vu le nombre de fois qu'on calcule le second membre de (1.31), on a rapporté cette quantité à M_k (fig. 1.1). On a cherché pour chaque M_k la solution w^{τ_k} du problème aux différences (1.32), (1.33) et calculé l'erreur

$$\zeta_k = \| w^{\tau_k} - u \|_{C,\ \tau_k}.$$

On définit w^{τ} à condition de résoudre au préalable le problème aux différences (1.5), (1.6), si bien que ζ_k a été rapporté à $2\, M_k$ (fig. 1.1).

On a regroupé dans le tableau 1.1, pour illustrer la convergence ponctuelle, les erreurs sur les solutions approchées relatives à plusieurs nœuds. En calcul automatique, les erreurs relatives d'arrondi se sont situées au niveau de 10^{-15}.

Tableau 1.1

t	Solution exacte $u(t)$	Erreur sur la solution du problème (1.5), (1.6)		Erreur sur la solution extrapolée (1.22) pour $\tau = 1/80$	Erreur sur la solution du système «douteux» (1.31) pour $\tau = 1/80$	Erreur sur la solution du problème (1.32), (1.33) pour $\tau = 1/80$
		$\tau = 1/80$	$\tau = 1/160$			
0,1	0,909090	$1,46 \cdot 10^{-7}$	$3,66 \cdot 10^{-8}$	$2,75 \cdot 10^{-11}$	$3,29 \cdot 10^{-6}$	$2,07 \cdot 10^{-8}$
0,2	0,833333	$4,48 \cdot 10^{-7}$	$1,20 \cdot 10^{-7}$	$4,12 \cdot 10^{-11}$	$6,93 \cdot 10^{-6}$	$1,88 \cdot 10^{-8}$
0,3	0,769230	$7,84 \cdot 10^{-7}$	$1,96 \cdot 10^{-7}$	$4,77 \cdot 10^{-11}$	$9,29 \cdot 10^{-6}$	$1,72 \cdot 10^{-8}$
0,4	0,714285	$1,10 \cdot 10^{-6}$	$2,75 \cdot 10^{-7}$	$5,03 \cdot 10^{-11}$	$1,18 \cdot 10^{-5}$	$1,56 \cdot 10^{-8}$
0,5	0,666666	$1,37 \cdot 10^{-6}$	$3,43 \cdot 10^{-7}$	$5,04 \cdot 10^{-11}$	$1,33 \cdot 10^{-5}$	$1,42 \cdot 10^{-8}$
0,6	0,625000	$1,60 \cdot 10^{-6}$	$3,99 \cdot 10^{-7}$	$4,93 \cdot 10^{-11}$	$1,49 \cdot 10^{-5}$	$1,29 \cdot 10^{-8}$
0,7	0,588235	$1,77 \cdot 10^{-6}$	$4,43 \cdot 10^{-7}$	$4,73 \cdot 10^{-11}$	$1,62 \cdot 10^{-5}$	$1,17 \cdot 10^{-8}$
0,8	0,555555	$1,90 \cdot 10^{-6}$	$4,75 \cdot 10^{-7}$	$4,49 \cdot 10^{-11}$	$1,69 \cdot 10^{-5}$	$1,07 \cdot 10^{-8}$
0,9	0,526315	$1,99 \cdot 10^{-6}$	$4,97 \cdot 10^{-7}$	$4,22 \cdot 10^{-11}$	$1,77 \cdot 10^{-5}$	$9,68 \cdot 10^{-9}$
1,0	0,500000	$2,04 \cdot 10^{-6}$	$5,11 \cdot 10^{-7}$	$3,94 \cdot 10^{-11}$	$1,79 \cdot 10^{-5}$	$7,98 \cdot 10^{-9}$

Le tableau et la figure confirment les estimations théoriques obtenues antérieurement. Quant à la solution v^τ de (1.31), son peu de précision provient de la propriété, pour la matrice de ce système « douteux », d'être mal conditionnée. A la lumière de ces faits, on peut dire de la méthode des différences d'ordre supérieur qu'elle s'est présentée en effet sous forme de mise en œuvre simple et stable d'un schéma aux différences plutôt mauvais dont le seul avantage est son ordre d'approximation élevé.

1.2. Théorème du développement

Dans le paragraphe précédent, nous avons donné un exemple de raffinement de la solution approchée par l'extrapolation de Richardson pour un problème concret. La méthode s'appuyait sur le développement

$$u^\tau = u + \tau^2 v + \eta^\tau \quad \text{sur} \quad \bar{\omega}_\tau , \tag{2.1}$$

avec v une fonction indépendante de τ. On se propose d'établir sous forme abstraite, pour une vaste classe de problèmes, des conditions suffisantes d'existence de (2.1) dont les termes du second membre sont en nombre quelconque.

Soit, dans l'espace $\mathbf{R}^n$ de dimension n, $n \geqslant 1$, un domaine borné Ω. On désigne sa fermeture par $\bar{\Omega}$. Soit un problème de la physique mathématique:

$$\begin{aligned} Lu &= f \quad \text{dans} \quad \Omega, \\ lu &= g \quad \text{sur} \quad D. \end{aligned} \tag{2.2}$$

Ici L et l sont des opérateurs différentiels linéaires et D est la frontière (ou une portion de frontière) de Ω. Les fonctions f, g, u ont pour domaine de définition Ω, D et $\bar{\Omega}$ respectivement.

On introduit dans les espaces des fonctions ainsi définies les classes $M_k(\Omega)$, $N_k(D)$ et $P_k(\bar{\Omega})$ dépendant d'un paramètre entier $k \geqslant 0$.

Dans la suite de l'exposé, ces classes caractériseront de règle la régularité du second membre de l'équation, des valeurs à la frontière et de la solution même. Leur choix est essentiellement dicté pour chaque problème concerné par les résultats de possibilité connus. S'agissant des équations du type elliptipue, on a, par exemple, pris naturellement pour M_k, N_k et P_k les espaces respectifs $C^{k+2}(\bar{\Omega})$, $C^{k+2+\alpha}(\Gamma)$, $C^{k+2+\alpha}(\bar{\Omega})$ des fonctions höldériennes. L'existence et la dérivabilité des solutions des équations différentielles ordinaires sont le mieux étudiées pour C^m $[0, 1]$, si bien qu'il y a intérêt à interpréter M_k et P_k en termes de ces espaces. L'ensemble D est alors réduit à un ou deux points, et la classe N_k devient $\mathbf{R}$ ou $\mathbf{R}^2$.

De ce point de vue, la condition suivante est en général relative au domaine de définition du problème et aux coefficients des équations (2.2), et elle caractérise la possibilité du problème pour les seconds membres réguliers.

CONDITION A. *Etant donnés k entier arbitraire, $0 \leqslant k \leqslant m$, et deux fonctions $f \in M_k(\Omega)$, $g \in N_k(D)$ quelconques, le problème* (2.2) *possède une solution unique $u \in P_k(\bar{\Omega})$.*

La résolution numérique du problème commence par la construction du réseau $\bar{\Omega}_h \subset \bar{\Omega}$ de h variable qui peut être aussi petit qu'on le veut. On cherche la solution approchée dans l'espace des fonctions discrètes définies aux nœuds de $\bar{\Omega}_h$, et on remplace le problème différentiel par un système (algébrique) d'équations aux différences finies qui sont définies aux nœuds de certaines parties finies $\breve{\Omega}_h \subset \Omega$ et $D_h \subset D$. Les parties $\bar{\Omega}_h$, $\breve{\Omega}_h$ et D_h sont les analogues discrets des ensembles $\bar{\Omega}$, Ω et D respectivement. Le problème s'écrit donc

$$\begin{aligned} L_h u^h &= f \quad \text{sur} \quad \breve{\Omega}_h , \\ l_h u^h &= g \quad \text{sur} \quad D_h . \end{aligned} \tag{2.3}$$

Ici L_h et l_h sont des opérateurs algébriques linéaires et u^h est une fonction discrète qui approche aux nœuds de $\bar{\Omega}_h$ la solution u du problème différentiel initial. On munit les espaces vectoriels des fonctions discrètes définies sur $\bar{\Omega}_h$, $\breve{\Omega}_h$, D_h des normes respectives $\|\cdot\|_{\bar{\Omega}_h}$, $\|\cdot\|_{\breve{\Omega}_h}$, $\|\cdot\|_{D_h}$.

On énonce en termes de ces normes une condition caractéristique de la possibilité du problème aux différences (2.3) et de la stabilité de sa solution.

CONDITION B. *Soit ψ^h une fonction discrète. Si elle a pour domaine de définition l'ensemble $\bar{\Omega}_h$ et est solution du problème*

$$L_h \psi^h = f^h \quad sur \quad \check{\Omega}_h, \qquad l_h \psi^h = g^h \quad sur \quad D_h, \tag{2.4}$$

avec f^h, g^h des fonctions discrètes définies sur $\check{\Omega}_h$, D_h respectivement, alors on a l'estimation

$$\| \psi^h \|_{\bar{\Omega}_h} \leqslant c \left(\| f^h \|_{\check{\Omega}_h} + \| g^h \|_{D_h} \right). \tag{2.5}$$

On note que (2.5) entraîne l'unicité pour (2.3). En effet, ce problème admet une solution unique pour n'importe quel second membre si le problème homogène associé n'a pas d'autre solution à part la solution triviale. Or, c'est justement le cas du problème homogène

$$L_h \xi^h = 0 \quad \text{sur} \quad \check{\Omega}_h,$$

$$l_h \xi^h = 0 \quad \text{sur} \quad D_h,$$

car l'estimation (2.5) entraîne $\| \xi^h \|_{\bar{\Omega}_h} = 0$. D'où $\xi^h = 0$ sur $\bar{\Omega}_h$.

Voici une condition qui se rapporte directement au procédé d'approximation des opérateurs différentiels par des relations aux différences.

CONDITION C. *Il existe pour toute fonction $\varphi \in P_k (\bar{\Omega})$, $0 \leqslant k \leqslant m$, les développements**

$$L_h \varphi = L\varphi + \sum_{j=1}^{k} h^j a_j + \sigma^h \quad \text{sur} \quad \check{\Omega}_h, \qquad l_h \varphi = l\varphi + \sum_{j=1}^{k} h^j b_j + \rho^h \quad \text{sur} \quad D_h, \tag{2.6}$$

les fonctions a_j, b_j étant indépendantes de h, $a_j \in M_{k-j}(\Omega)$, $b_j \in N_{k-j}(D)$ et les restes σ^h, ρ^h étant évalués par

$$\| \sigma^h \|_{\check{\Omega}_h} \leqslant c_1 h^{k+\beta}, \qquad \| \rho^h \|_{D_h} \leqslant c_2 h^{k+\beta}, \tag{2.7}$$

où les constantes c_1, c_2 sont indépendantes de h et $\beta > 0$ ne dépend pas de h, k et φ.

Avec les conditions citées, on obtient pour la solution discrète u^h du problème (2.3) un développement analogue à (2.6).

---

* La somme dont la limite supérieure est plus petite que celle inférieure est considérée nulle, et le produit jouissant de la même propriété est 1.

THÉORÈME 2.1. *On suppose que les problèmes* (2.2) *et* (2.3) *vérifient les conditions* A, B *et* C *et que* $f \in M_m(\Omega)$, $g \in N_m(D)$. *La solution discrète* u^h *admet le développement*

$$u^h = u + \sum_{j=1}^{m} h^j v_j + \eta^k \quad sur \quad \bar{\Omega}_h. \tag{2.8}$$

Ici les fonctions v_j *sont indépendantes de* h, $v_j \in P_{m-j}(\bar{\Omega})$, *et le reste* η^h *els majoré par*

$$\| \eta^h \|_{\bar{\Omega}_h} \leqslant c_3 h^{m+\beta}, \tag{2.9}$$

la constante c_3 *étant indépendante de* h.

DÉMONSTRATION. Soit un ensemble quelconque de fonctions $v_j \in P_{m-j}(\bar{\Omega})$, $j = 1, \ldots, m$, indépendantes de h. Connaissant v_j et deux solutions u et u^k, on définit la fonction discrète

$$\eta^k = u^h - u - \sum_{j=1}^{m} h^j v_j \quad \text{sur} \quad \bar{\Omega}_h. \tag{2.10}$$

On porte dans (2.3) u^h exprimée à partir de (2.10), il vient

$$\begin{aligned} L_h u + \sum_{j=1}^{m} h^j L_h v_j + L_h \eta^k &= f \quad \text{sur} \quad \check{\Omega}_h, \\ l_h u + \sum_{j=1}^{m} h^j l_h v_j + l_h \eta^h &= g \quad \text{sur} \quad D_h. \end{aligned} \tag{2.11}$$

On a, conformément à la condition C,

$$\begin{aligned} L_h u &= f + \sum_{i=1}^{m} h^i a_{0,i} + \sigma_0^h \quad \text{sur} \quad \check{\Omega}_h, \\ l_h u &= g + \sum_{i=1}^{m} h^i b_{0,i} + \rho_0^h \quad \text{sur} \quad D_h \end{aligned} \tag{2.12}$$

et

$$\begin{aligned} L_h v_j &= L v_j + \sum_{i=1}^{m-j} h^i a_{j,i} + \sigma_j^h \quad \text{sur} \quad \check{\Omega}_h, \\ l_h v_j &= l v_j + \sum_{i=1}^{m-j} h^i b_{j,i} + \rho_j^h \quad \text{sur} \quad D_h. \end{aligned} \tag{2.13}$$

Ici

$$a_{j,i} \in M_{m-j-i}(\Omega), \qquad b_{j,i} \in N_{m-j-i}(D), \tag{2.14}$$

$a_{j,i}$, $b_{j,i}$ ne dépendent pas de h, et les restes vérifient les inégalités

$$\| \sigma_j^h \|_{\check{\Omega}_h} \leqslant c_{j,1}\, h^{m-j+\beta}, \qquad \| \rho_j^h \|_{D_h} \leqslant c_{j,2}\, h^{m-j+\beta}, \tag{2.15}$$

avec les constantes $c_{j,1}$ et $c_{j,2}$ indépendantes de h. On se sert des développements (2.12), (2.13) pour ramener (2.11) à la forme

$$\begin{aligned} &f + \sum_{j=1}^{m} h^j L v_j + \sum_{j=0}^{m} h^j \sum_{i=1}^{m-j} h^i a_{j,i} + \sum_{j=0}^{m} h^j \sigma_j^h + L_h \eta^h = f \quad \text{sur} \quad \check{\Omega}_h, \\ &g + \sum_{j=1}^{m} h^j l v_j + \sum_{j=0}^{m} h^j \sum_{i=1}^{m-j} h^i b_{j,i} + \sum_{j=0}^{m} h^j \rho_j^h + l_h \eta^h = g \quad \text{sur} \quad D_h. \end{aligned} \tag{2.16}$$

On pose

$$\xi^h = \sum_{j=0}^{m} h^j \sigma_j^h, \qquad \zeta^h = \sum_{j=0}^{m} h^j \rho_j^h$$

et on utilise les estimations (2.15), il vient

$$\| \xi^h \|_{\check{\Omega}_h} \leqslant h^{m+\beta} c_4, \qquad \| \zeta^h \|_{D_h} \leqslant h^{m+\beta}\, c_5, \tag{2.17}$$

où

$$c_4 = \sum_{j=0}^{m} c_{j,1}, \qquad c_5 = \sum_{j=0}^{m} c_{j,2}.$$

Avec les notations introduites, les relations (2.16) deviennent au prix de plusieurs transformations simples

$$\begin{aligned} &\sum_{j=1}^{m} h^j \Big(L v_j + \sum_{i=1}^{j} a_{j-i,i} \Big) + \xi^h + L_h \eta^h = 0 \quad \text{sur} \quad \check{\Omega}_h, \\ &\sum_{j=1}^{m} h^j \Big(l v_j + \sum_{i=1}^{j} b_{j-i,i} \Big) + \zeta^h + l_h \eta^h = 0 \quad \text{sur} \quad D_h. \end{aligned} \tag{2.18}$$

Ainsi, on a obtenu, pour un ensemble quelconque de fonctions $v_j \in P_{m-j}(\Omega)$ et pour η^h définie par (2.10), les égalités (2.18) dont les restes ξ^h et ζ^h admettent les majorations (2.17).

On assimile maintenant v_j, $j = 1, 2, \ldots, m$, aux solutions des problèmes différentiels

$$\begin{aligned} L v_j &= - \sum_{i=1}^{j} a_{j-i,i} \quad \text{dans} \quad \Omega, \\ l v_j &= - \sum_{i=1}^{j} b_{j-i,i} \quad \text{sur} \quad D. \end{aligned} \tag{2.19}$$

La fonction v_1 est par exemple solution du problème

$$Lv_1 = -a_{0,1} \quad \text{dans} \quad \Omega,$$
$$lv_1 = -b_{0,1} \quad \text{sur} \quad D.$$

La condition C entraîne pour le développement (2.12) que $a_{0,1} \in M_{m-1}(\Omega)$ et $b_{0,1} \in N_{m-1}(D)$. Aussi v_1 est définie univoquement et $v_1 \in P_{m-1}(\bar{\Omega})$ (voir condition A). On suppose connues $v_j \in P_{m-j}(\bar{\Omega})$ pour $j=1, \ldots, k$, $1 \leqslant k \leqslant m$. La condition C implique pour $j = 1, \ldots, k$ la validité de k développements (2.13) vérifiant (2.14). On écrit le problème (2.19) pour $j=k+1$:

$$\begin{aligned} Lv_{k+1} &= -\sum_{i=1}^{k+1} a_{k-i+1,\, i} \quad \text{dans} \quad \Omega, \\ lv_{k+1} &= -\sum_{i=1}^{k+1} b_{k-i+1,\, i} \quad \text{sur} \quad D. \end{aligned} \tag{2.20}$$

Etant donné (2.14), les seconds membres sont dans $M_{m-k-1}(\Omega)$ et $N_{m-k-1}(D)$ respectivement, si bien que le problème (2.20) possède, par suite de la condition A, une solution unique $v_{k+1} \in P_{m-k-1}(\bar{\Omega})$. Cette solution est manifestement indépendante de h.

Ainsi, nous avons donné pour $j=1, \ldots, m$ un procédé de génération des fonctions $v_j \in P_{m-j}(\bar{\Omega})$ indépendantes de h. Cette collection de v_j satisfait également aux identités (2.18) et aux estimations (2.17), et les relations (2.18) s'écrivent en raison de (2.19):

$$L_h \eta^h = -\xi^h \quad \text{sur} \quad \bar{\Omega}_h,$$
$$l_h \eta^h = -\zeta^h \quad \text{sur} \quad D_h.$$

La condition B détermine l'inégalité

$$\| \eta^h \|_{\bar{\Omega}_h} \leqslant c \left(\| \xi^h \|_{\bar{\Omega}_h} + \| \zeta^h \|_{D_h} \right).$$

Avec les majorations (2.17), on obtient (2.9) où $c_3 = c(c_4 + c_5)$. On exprime u^h moyennant (2.10) et on est conduit à l'affirmation du théorème, i.e. au développement (2.8) ayant les propriétés voulues.

Dans le paragraphe suivant, nous analyserons une méthode d'amélioration des solutions aux différences qui s'appuie sur le développement démontré, et nous reprenons pour le moment la condition C. Nombreux sont les problèmes aux différences où les coefficients a_j, b_j de (2.6) s'annulent pour j impairs. Un exemple en a été donné dans le paragraphe précédent. Par conséquent, nous sommes amenés au développement (2.8) suivant les puissances paires seules de h. Cette circonstance permettra de réduire sensiblement le nombre d'opérations en calcul automatique. On énonce donc

la variante correspondante de la condition C et le théorème du développement associé.

CONDITION D. *Il existe pour toute fonction* $\varphi \in P_{m-2k}(\bar{\Omega})$, $k = 0, 1, \ldots, s$, $s = [m/2]$, *les développements*

$$\begin{aligned} L_h \varphi &= L\varphi + \sum_{j=1}^{s-k} h^{2j} a_j + \sigma^h \quad sur \quad \breve{\Omega}_h. \\ l_h \varphi &= l\varphi + \sum_{j=1}^{s-k} h^{2j} b_j + \rho^h \quad sur \quad D_h. \end{aligned} \tag{2.21}$$

les fonctions a_j, b_j *étant indépendantes de* h, $a_j \in M_{m-2k-2j}(\Omega)$, $b_j \in N_{m-2k-2j}(D)$ *et les restes étant évalués par*

$$\| \sigma^h \|_{\breve{\Omega}_h} \leqslant c_6\, h^{m-2k+\beta}, \quad \| \rho^h \|_{D_h} \leqslant c_7\, h^{m-2k+\beta}, \tag{2.22}$$

ou les constantes c_6 c_7 *sont indépendantes de* h *et* β *ne dépend pas de* h, k *et* φ.

THÉORÈME 2.2 *On suppose que le problème différentiel* (2.2), *et son analogue aux différences finies* (2.3) *vérifient les conditions* A, B *et* D *et que* $f \in M_m(\Omega)$, $g \in N_m(D)$. *La solution discrète* u^h *admet le développement*

$$u^h = u + \sum_{j=1}^{s} h^{2j} v_j + \eta^h \quad sur \quad \bar{\Omega}_h. \tag{2.23}$$

Ici les fonctions v_j *sont indépendantes de* h, $v_j \in P_{m-2j}(\bar{\Omega})$, *et le reste* η^h *est majoré par*

$$\| \eta^h \|_{\bar{\Omega}_h} \leqslant c_8\, h^{m+\beta}. \tag{2.24}$$

la constante c^2 *étant indépendante de* h.

La démonstration des relations (2.23), (2.24) est analogue à celle du théorème précédent à la différence que tous les calculs relatifs à la partie régulière des développements ne s'effectuent qu'avec les puissances paires de h.

REMARQUE. Dans plusieurs problèmes, l'équation aux différences $l_h u^h = g$ coïncide avec la condition aux limites $lu = g$ sur D_h, si bien que b_j et ρ^h de (2.6) et (2.21) s'annulent. Alors on simplifie la condition B sans altérer les résultats des théorèmes 2.1 et 2.2, à savoir on a la

CONDITION B'. *On suppose que la fonction discrète* ψ^h *est définie sur* $\bar{\Omega}_h$ *et est solution du problème*

$$\begin{aligned} L_h \psi^h &= f^h \quad sur \quad \breve{\Omega}_h. \\ l_h \psi^h &= 0 \quad sur \quad D_h. \end{aligned} \tag{2.25}$$

f^h étant une fonction discrète de domaine de définition $\breve{\Omega}_h$. Alors ψ^h est majorée par

$$\| \psi^h \|_{\bar{\Omega}_h} \leqslant c \, \| f^h \|_{\breve{\Omega}_h}. \tag{2.26}$$

Cette condition entraîne elle aussi l'unicité de la solution du problème discret (2.3).

1.3. Accélération de la convergence

Nous allons utiliser les développements des théorèmes 2.1 et 2.2 pour raffiner les solutions approchées du problème (2.3). On munit les fonctions discrètes définies sur $\bar{\Omega}_h$ d'une norme uniforme, i.e.

$$\| v \|_{\bar{\Omega}_h} = \max_{x \in \bar{\Omega}_h} | v(x) |. \tag{3.1}$$

On étudiera d'autres cas au fur et à mesure qu'ils interviendront.

On suppose qu'une suite de domaines de discrétisation $\bar{\Omega}_{h_k}$ de pas $h_1 > h_2 > \ldots > h_{m+1} > 0$ vérifient les conditions du théorème 2.1. On exige que ces réseaux possèdent l'intersection non vide:

$$\bar{\Omega}_H = \bigcap_{k=1}^{m+1} \bar{\Omega}_{h_k} \neq \varnothing.$$

Selon la condition B, le problème aux différences finies

$$\left. \begin{aligned} L_h u^h &= f \quad \text{sur} \quad \breve{\Omega}_h, \\ l_h u^h &= g \quad \text{sur} \quad D_h \end{aligned} \right. \tag{3.2}$$

admet une solution unique pour chaque valeur du paramètre $h = h_k$. On désigne cette solution par u^{h_k}. Toutes les u^{h_k} sont définies sur $\bar{\Omega}_H$.

Considérons le système

$$\begin{aligned} \sum_{k=1}^{m+1} \gamma_k &= 1, \\ \sum_{k=1}^{m+1} \gamma_k h_k^j &= 0, \qquad j = 1, \ldots, m. \end{aligned} \tag{3.3}$$

Si l'on lit le début du § 7.2, on voit que son déterminant est $\neq 0$, si bien qu'il y a unicité. On forme la combinaison linéaire avec les poids connus γ_k:

$$U^H(x) = \sum_{k=1}^{m+1} \gamma_k u^{h_k}(x), \qquad x \in \bar{\Omega}_H, \tag{3.4}$$

et on démontre que la solution U^H est plus précise que chacune des solutions u^{h_k}.

THÉORÈME 3.1. *On suppose que les domaines discrétisés* $\bar{\Omega}_{h_k}$ *de paramètres* $h_1 > \ldots > h_m > h_{m+1} > 0$ *satisfont aux conditions du théorème* 2.1 *pour la norme uniforme* (3.1) *et qu'on a l'inégalité*

$$h_k/h_{k+1} \geqslant 1 + d_1, \qquad k = 1, \ldots, m, \tag{3.5}$$

avec la constante $d_1 > 0$ *indépendante de* h_k. *On a dans l'intersection non vide* $\bar{\Omega}_H$ *l'estimation*

$$\max_{\bar{\Omega}_H} | U^H - u | \leqslant d_2 \, h_1^{m+\beta}, \tag{3.6}$$

où U^H *est la solution extrapolée* (3.4) *munie des poids* γ_k *définis à partir du système* (3.3), u *la solution du problème différentiel* (2.2) *et* d_2 *une constante indépendante de* h_k.

DÉMONSTRATION. On fixe un point quelconque x de $\bar{\Omega}_H$. D'après le théorème 2.1, on a en ce point les développements

$$u^{h_k}(x) = u(x) + \sum_{j=1}^{m} h_k^j \, v_j(x) + \eta^{h_k}(x), \qquad k = 1, 2, \ldots, m+1, \tag{3.7}$$

avec $v_j(x)$ indépendantes de h_k et les restes admettant la majoration

$$| \eta^{h_k}(x) | \leqslant \| \eta^{h_k} \|_{\bar{\Omega}_{h_k}} \leqslant c_3 \, h_k^{m+\beta}. \tag{3.8}$$

On transforme le second membre de (3.4) à l'aide du développement (3.7), il vient

$$U^H(x) = \sum_{k=1}^{m+1} \gamma_k u(x) + \sum_{k=1}^{m+1} \sum_{j=1}^{m} \gamma_k \, h_k^j v_j(x) + \sum_{k=1}^{m+1} \gamma_k \, \eta^{h_k}(x). \tag{3.9}$$

Les quantités $v_j(x)$ étant indépendantes de k, le système (3.3) entraîne

$$\sum_{k=1}^{m+1} \sum_{j=1}^{m} \gamma_k \, h_k^j v_j(x) = \sum_{j=1}^{m} v_j(x) \sum_{k=1}^{m+1} \gamma_k \, h_k^j = 0.$$

De plus,

$$\sum_{k=1}^{m+1} \gamma_k \, u(x) = u(x).$$

Aussi l'égalité (3.9) se récrit

$$U^H(x) = u(x) + \sum_{k=1}^{m+1} \gamma_k \, \eta^{h_k}(x).$$

D'où

$$| U^H(x) - u(x) | \leqslant \sum_{k=1}^{m+1} | \gamma_k | \, | \eta^{h_k}(x) | . \qquad (3.10)$$

On évalue $| \gamma_k |$ par recours au lemme 2.3, § 7.2. Ce lemme implique en vertu de (3.5):

$$| \gamma_k | \leqslant \left(\frac{1+d_1}{d_1}\right)^m \cdot \qquad k = 1, \ldots, m+1.$$

Avec cette estimation et l'inégalité (3.8), on ramène (3.10) à

$$| U^H(x) - u(x) | \leqslant \sum_{k=1}^{m+1} \left(\frac{1+d_1}{d_1}\right)^m c_3 h_k^{m+\beta} \leqslant c_3 m \left(\frac{1+d_1}{d_1}\right)^m h_1^{m+\beta}.$$

On pose

$$d_2 = c_3 m \left(\frac{1+d_1}{d_1}\right)^m,$$

auquel cas

$$| U^H(x) - u(x) | \leqslant d_2 h_1^{m+\beta}, \qquad x \in \bar{\Omega}_H.$$

d'où (3.6), i.e. le résultat cherché.

Voyons deux procédés de resserrement des réseaux parmi les plus répandus. Le premier consiste à choisir une succession de domaines discrétisés $\bar{\Omega}_{h_k}$ associés aux paramètres $h_k = h/k$, où $h > 0$, $k = 1, \ldots, m+1$. Dans ce cas, la condition (3.5) est remplie pour tout $h > 0$, et la constante $d_1 = 1/m$. Le système (3.3) devient pour ces paramètres

$$\sum_{k=1}^{m+1} \gamma_k = 1,$$

$$\sum_{k=1}^{m+1} \frac{\gamma_k}{k^j} = 0, \qquad j = 1, \ldots, m.$$

Conformément au lemme 2.1, § 7.2, sa solution est cherchée sous forme explicite

$$\gamma_k = \frac{(-1)^{m-k+1} k^{m+1}}{k!\,(m-k+1)!}, \qquad k = 1, \ldots, m+1. \qquad (3.11)$$

Le tableau ci-dessous donne les poids γ_k calculés par (3.11) pour plusieurs m.

Tableau 1.2

m	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
1	-1	2				
2	$\frac{1}{2}$	-4	$\frac{9}{2}$			
3	$-\frac{1}{6}$	4	$-\frac{27}{2}$	$\frac{32}{3}$		
4	$\frac{1}{24}$	$-\frac{8}{3}$	$\frac{81}{4}$	$-\frac{128}{3}$	$\frac{625}{24}$	
5	$-\frac{1}{120}$	$\frac{4}{3}$	$-\frac{81}{4}$	$\frac{256}{3}$	$-\frac{3125}{24}$	$\frac{324}{5}$
Paramètre	h	$h/2$	$h/3$	$h/4$	$h/5$	$h/6$

On voit que γ_k croissent sensiblement avec l'augmentation de m, ce qui entraîne l'importance plus grande des erreurs d'arrondi et des autres erreurs irrégulières dues à l'imprécision de la solution des problèmes aux différences (3.2). Si l'on travaille avec une capacité faible, ce facteur risque de s'avérer décisif. Dans le second procédé de resserrement, la croissance des poids γ_k est moins catastrophique.

On prend $h_k = h/2^{k-1}$, où $h > 0$, $k = 1, \ldots, m+1$, auquel cas la condition (3.5) est juste pour tout $h > 0$ et la constante $d_1 = 1$. Le système (3.3) se récrit

$$\sum_{k=1}^{m+1} \gamma_k = 1,$$
$$\sum_{k=1}^{m+1} \frac{\gamma_k}{2^{j(k-1)}} = 0, \qquad j = 1, \ldots, m. \tag{3.12}$$

Ses solutions associées à plusieurs m sont données dans le tableau 1.3.

On voit sans peine que γ_k augmentent plus lentement.

S'agissant des développements (2.8) à partie régulière contenant de nombreux termes, le dernier choix des paramètres h_k garantit des résultats plus précis. La résolution des problèmes (3.2) avec h, $h/2$, $h/4$, ... exige par contre une masse de calculs sensiblement plus importante que dans le premier cas. Diminuer le pas de moitié équivaut de règle pour les équations différentielles ordinaires à

Tableau 1.3

m	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
1	-1	2				
2	$\frac{1}{3}$	-2	$\frac{8}{3}$			
3	$-\frac{1}{21}$	$\frac{2}{3}$	$-\frac{8}{3}$	$\frac{64}{21}$		
4	$\frac{1}{315}$	$-\frac{2}{21}$	$\frac{8}{9}$	$-\frac{64}{21}$	$\frac{1024}{315}$	
5	$-\frac{1}{9765}$	$\frac{2}{315}$	$-\frac{8}{63}$	$\frac{64}{63}$	$-\frac{1024}{315}$	$\frac{32768}{9765}$
Paramètre	h	$h/2$	$h/4$	$h/8$	$h/16$	$h/32$

augmenter de deux fois l'effort de calcul (cette augmentation est encore plus grande pour les problèmes à plusieurs variables).

Il y a donc intérêt à prendre plusieurs paramètres successifs de la série (voir [131])

$$h,\ h/2,\ h/3,\ h/4,\ h/6,\ h/8,\ h/12,\ \ldots$$

pour lesquels la quantité de calcul à exécuter ne croît pas trop avec la diminution des pas. D'autre part, la constante d_1 de la condition (3.5) garantissant la propriété de borne de $|\gamma_k|$ est prise égale à 1/3 quelle que soit la longueur de la série partielle choisie.

On note qu'on calcule la somme

$$\sum_{k=1}^{m+1} \gamma_k\, u^{h_k}(x)$$

par l'algorithme de Neville sans résoudre au préalable le système (3.3). En effet, soit le tableau d'extrapolation

$$\begin{array}{llll}
T_1^{(0)} \rightarrow & T_1^{(1)} \ldots & T_1^{(m-1)} & T_1^{(m)} \\
T_2^{(0)} \nearrow & T_2^{(1)} & T_2^{(m-1)} & \\
\vdots & \vdots \quad \cdot^{\cdot^{\cdot}} & & \\
T_m^{(0)} & T_m^{(1)} & & \\
T_{m+1}^{(0)} & & &
\end{array}$$

On pose

$$T_j^{(0)} = u^{h_j}(x), \qquad j = 1, 2, \ldots, m+1.$$

On procède par itérations en calculant de proche en proche les éléments des colonnes de numéros $i = 1, 2, \ldots, m$:

$$T_j^{(i)} = \frac{h_{i+j}\, T_j^{(i-1)} - h_j\, T_{j+1}^{(i-1)}}{h_{i+j} - h_j}, \qquad j = 1, \ldots, m - i + 1.$$

On obtient finalement

$$T_1^{(m)} = \sum_{k=1}^{m+1} \gamma_k\, u^{h_k}(x).$$

La démonstration de cette égalité est faite au § 7.2.

L'algorithme de Neville est particulièrement indiqué pour m grands si le calcul direct de γ_k à partir du système (3.3) s'avère difficile ou si l'ordre m de l'extrapolation de Richardson est choisi au cours du calcul. Dans le cas particulier des réseaux de h tels que

$$h_i = h/a^{i-1}, \quad \text{où} \quad h > 0, \quad a > 1, \quad i = 1, 2, \ldots, m+1,$$

l'algorithme de Neville conduit à la règle de Romberg (voir [125]).

On énonce des résultats analogues pour le cas important où la partie régulière du développement ne renferme que les puissances paires de h.

Soit m un entier naturel et $s = [m/2]$. On construit les domaines discrétisés $\bar{\Omega}_{h_k}$ avec les paramètres $h_1 > \ldots > h_{s+1} > 0$ et on suppose remplies les conditions du théorème 2.2. Alors on définit sur chaque réseau $\bar{\Omega}_{h_k}$ une solution u^{h_k} du problème aux différences (3.2). Si

$$\bar{\Omega}_H = \bigcap_{k=1}^{s+1} \bar{\Omega}_{h_k} \neq \varnothing,$$

alors toutes les solutions sont définies sur $\bar{\Omega}_H$.

Soit le système

$$\begin{aligned} &\sum_{k=1}^{s+1} \gamma_k = 1, \\ &\sum_{k=1}^{s+1} \gamma_k\, h_k^{2j} = 0, \qquad j = 1, \ldots, s. \end{aligned} \tag{3.13}$$

Comme $h_k \neq h_j$, $k \neq j$, les résultats du § 7.2 entraînent que le déterminant du système est non nul. Il existe donc une seule

solution $\gamma_1, \ldots, \gamma_{s+1}$. On forme la combinaison linéaire avec les poids γ_k:

$$U^H(x) = \sum_{k=1}^{s+1} \gamma_k u^{h_k}(x), \qquad x \in \bar{\Omega}_H. \tag{3.14}$$

et on démontre que U^H approche u mieux que u^{h_k}.

Théorème 3.2. *Etant donnés les paramètres $h_1 > \ldots > h_{s+1} > 0$ tels que*

$$h_k/h_{k+1} \geq 1 + d_3, \qquad k = 1, \ldots, s, \tag{3.15}$$

avec la constante $d_3 > 0$ indépendante de h_k, on suppose que le problème (3.2) *et les réseaux $\bar{\Omega}_{h_k}$ vérifient les conditions du théorème* 2.2 *pour la norme uniforme* (3.1). *Dans l'intersection $\bar{\Omega}_H$ des réseaux a lieu l'estimation*

$$\max_{\bar{\Omega}_H} |U^H - u| \leq d_4 h_1^{m+\beta}, \tag{3.16}$$

où U^H est la solution extrapolée (3.14) *munie des poids γ_k définis à partir du système* (3.13), *u la solution du problème différentiel* (2.2) *et d_4 une constante indépendante de h_k.*

Démonstration. On fixe x quelconque de $\bar{\Omega}_H$. En vertu du théorème 2.2, on a en ce point les développements

$$u^{h_k}(x) = u(x) + \sum_{j=1}^{s} h_k^{2j} v_j(x) + \eta^{h_k}(x), \qquad k = 1, \ldots, s+1.$$

On les porte dans le second membre de (3.14), il vient

$$U^H(x) = \sum_{k=1}^{s+1} \gamma_k u(x) + \sum_{k=1}^{s+1} \sum_{j=1}^{s} h_k^{2j} \gamma_k v_j(x) + \sum_{k=1}^{s+1} \gamma_k \eta^{h_k}(x). \tag{3.17}$$

Les quantités $v_j(x)$ ne dépendent pas de k, et γ_k satisfont au système (3.13), si bien que

$$\sum_{k=1}^{s+1} \sum_{j=1}^{s} h_k^{2j} \gamma_k v_j(x) = \sum_{j=1}^{s} v_j(x) \sum_{k=1}^{s+1} \gamma_k h_k^{2j} = 0.$$

La première équation du système (3.13) entraîne

$$\sum_{k=1}^{s+1} \gamma_k u(x) = u(x);$$

c'est pourquoi

$$U^H(x) = u(x) + \sum_{k=1}^{s+1} \gamma_k \eta^{h_k}(x).$$

D'où

$$| U^H(x) - u(x) | \leqslant \sum_{k=1}^{s+1} | \gamma_k | | \eta^{h_k}(x) | . \qquad (3.18)$$

Les restes η^{h_k} sont évalués par

$$| \eta^{h_k}(x) | \leqslant c_8 h_k^{m+\beta}$$

(th. 2.2). Etant donné (3.15), on est pour (3.13) dans les conditions du lemme 2.4, § 7.2, ce qui permet d'évaluer les poids γ_k :

$$| \gamma_k | \leqslant \left(\frac{1 + 2d_3 + d_3^2}{2d_3 + d_3^2} \right)^s . \qquad k = 1, \ldots, s + 1.$$

Avec les estimations obtenues, l'inégalité (3.18) devient

$$| U^H(x) - u(x) | \leqslant \sum_{k=1}^{s+1} \left(\frac{1 + 2d_3 + d_3^2}{2d_3 + d_3^2} \right)^s c_8 h_k^{m+\beta} \leqslant$$

$$\leqslant c_8 \left(\frac{1 + 2d_3 + d_3^2}{2d_3 + d_3^2} \right)^s h_1^{m+\beta} .$$

On pose

$$d_4 = c_8 \left(\frac{1 + 2d_3 + d_3^2}{2d_3 + d_3^2} \right)^s .$$

Comme d_4 est indépendant de h_k, on a

$$| U^H(x) - u(x) | \leqslant d_4 h_1^{m+\beta} \qquad \forall x \in \bar{\Omega}_H,$$

i.e. le résultat cherché.

Quel sera le comportement des solutions du système (3.13) si l'on resserre les nœuds des réseaux par deux procédés décrits plus haut? Dans le premier cas, on a choisi les paramètres par la formule $h_k = h/k$, où $h > 0$, $k = 1, \ldots, s + 1$. Il est aisé de vérifier la condition (3.15) : quel que soit $h > 0$, la constante d_3 est égale à $1/s$. Avec ce jeu de paramètres, le système (3.13) s'écrit

$$\sum_{k=1}^{s+1} \gamma_k = 1,$$

$$\sum_{k=1}^{s+1} \frac{\gamma_k}{k^{2j}} = 0, \qquad j = 1, \ldots, s.$$

Le lemme 2.2, § 7.2, implique la formule

$$\gamma_k = 2 \frac{(-1)^{s-k+1} \, k^{2s+2}}{(s+k+1)!\,(s-k+1)!}, \qquad k = 1, \ldots, s+1.$$

Le tableau 1.4 donne pour plusieurs m les valeurs des poids ainsi calculés.

Tableau 1.

m	s	γ_1	γ_2	γ_3	γ_4	γ_5
2 3	1	$-\frac{1}{3}$	$\frac{4}{3}$			
4 5	2	$\frac{1}{24}$	$-\frac{16}{15}$	$\frac{81}{40}$		
6 7	3	$-\frac{1}{360}$	$\frac{16}{45}$	$-\frac{729}{280}$	$\frac{1024}{315}$	
8 9	4	$\frac{1}{8640}$	$-\frac{64}{945}$	$\frac{6561}{4480}$	$-\frac{16384}{2835}$	$\frac{390625}{72576}$
Paramètre		h	$h/2$	$h/3$	$h/4$	$h/5$

On signale qu'avec l'augmentation de m, γ_k croissent en valeur absolue plus lentement que dans le cas général où la partie régulière du développement renferme des puissances impaires de h.

Le second procédé a consisté à choisir h_k tels que $h_k = h/2^{k-1}$, $k = 1, \ldots, s+1$, avec $h > 0$ une valeur initiale. On établit sans peine que la constante d_3 de (3.15) peut être prise égale à 1. Le système (3.13) devient par des transformations insignifiantes

$$\begin{aligned} &\sum_{k=1}^{s+1} \gamma_k = 1, \\ &\sum_{k=1}^{s+1} \frac{\gamma_k}{2^{2j(k-1)}} = 0, \qquad j = 1, \ldots, s. \end{aligned} \tag{3.19}$$

Voici ses solutions pour plusieurs m (voir tableau 1.5).

La croissance en valeur absolue des γ_k est encore plus tempérée. S'agissant de m importants, il y a intérêt à renoncer au calcul direct de la somme

$$U^H(x) = \sum_{k=1}^{s+1} \gamma_k \, u^{h_k}(x)$$

Tableau 1.5

m	s	γ_1	γ_2	γ_3	γ_4	γ_5
2 3	1	$-\frac{1}{3}$	$\frac{4}{3}$			
4 5	2	$\frac{1}{45}$	$-\frac{4}{9}$	$\frac{64}{45}$		
6 7	3	$-\frac{1}{2835}$	$\frac{4}{135}$	$-\frac{64}{135}$	$\frac{4096}{2835}$	
8 9	4	$\frac{1}{722925}$	$-\frac{4}{8505}$	$\frac{64}{2025}$	$-\frac{4096}{8505}$	$\frac{1048576}{722925}$
Paramètre		h	$h/2$	$h/4$	$h/8$	$h/16$

avec les poids trouvés à partir de (3.13) au profit d'une variante de l'algorithme de Neville. On reprend le tableau d'extrapolation de la page 36 et on a pour la colonne 0:

$$T_j^{(0)} = u^{h_j}(x), \qquad j = 1, 2, \ldots, m+1. \tag{3.20}$$

Les colonnes suivantes sont calculées par la formule récurrentielle

$$T_j^{(i)} = \frac{h_{i+j}^2\, T_j^{(i-1)} - h_j^2\, T_{j+1}^{(i-1)}}{h_{i+j}^2 - h_j^2}, \qquad j = 1, \ldots, m-i+1; \quad i = 1, 2, \ldots, m. \tag{3.21}$$

Finalement,

$$T_1^{(m)} = \sum_{k=1}^{m} \gamma_k\, u^{h_k}(x)$$

(pour la démonstration voir § 7.2).

1.4. Raffinement par les différences d'ordre supérieur

Soit, en notations du § 1.2, le problème approché d'ordre élevé

$$S_h u^h = f \quad \text{sur} \quad \Omega_h, \qquad s_h u^h = g \quad \text{sur} \quad D_h, \tag{4.1}$$

où S_h, s_h sont des opérateurs algébriques linéaires qui sont des approximations d'ordre élevé des opérateurs différentiels L, l sur les

réseaux $\bar{\Omega}_h$, D_h. Ils sont de règle plus compliqués que L_h, l_h. On suppose qu'on est pour ces opérateurs dans la

CONDITION E. *Toute fonction* $\varphi \in P_k(\bar{\Omega})$, $0 \leqslant k \leqslant m$, *vérifie les inégalités*

$$\| S_h \varphi - L \varphi \|_{\bar{\Omega}_h} \leqslant c_9 h^{k+\beta},$$
$$\| s_h \varphi - l \varphi \|_{D_h} \leqslant c_{10} h^{k+\beta}.$$

Si l'on démontre la condition B pour le problème (4.1), il existe une solution unique u^h. Selon le théorème de convergence (voir p. ex. [112]), la stabilité et l'approximation entraînent la convergence de u^h vers la solution exacte, et cette convergence est d'ordre élevé. Si l'on est, disons, dans la condition de régularité $u \in P_m(\bar{\Omega})$, il en résulte l'estimation

$$\| u^h - u \|_{\bar{\Omega}_h} \leqslant c_{11} h^{m+\beta}. \tag{4.2}$$

La résolution du problème (4.1) s'avère délicate si sa matrice est de structure compliquée. Si la stabilité de (4.1) n'est pas justifiée, toute tentative de le résoudre reste pratiquement vaine par suite du conditionnement mauvais ou de la dégénérescence de sa matrice.

Si la recherche de la solution du problème (2.3) est de règle plus simple, il approche par contre le problème différentiel à l'ordre peu élevé. Considérons dans cette optique plusieurs itérations du procédé

$$L^h u^h_{k+1} = f + L^h u^h_k - S^h u^h_k \quad \text{sur} \quad \bar{\Omega}_h, \tag{4.3}$$
$$l^h u^h_{k+1} = g + l^h u^h_k - s^h u^h_k \quad \text{sur} \quad D_h. \tag{4.4}$$
$$k = 0, 1, \ldots;$$
$$u^h_0 = 0 \quad \text{sur} \quad \bar{\Omega}_h. \tag{4.5}$$

On rèsout à chaque pas un problème de la forme (2.3). On montre que la solution obtenue u^h_k approche sous certaines conditions la fonction u avec une grande précision. On a besoin de la

CONDITION F. *La solution du problème*

$$L^h w^h = L^h v^h - S^h v^h \quad sur \quad \bar{\Omega}_h,$$
$$l^h w^h = l^h v^h - s^h v^h \quad sur \quad D_h,$$

où v^h *est une fonction discrète quelconque définie sur* $\bar{\Omega}_h$, *admet la majoration*

$$\| w^h \|_{\bar{\Omega}_h} \leqslant c_{12} \| v^h \|_{\bar{\Omega}_h}.$$

Les conditions citées sont suffisantes pour justifier une précision toujours plus grande des u_k^h successives obtenues par itérations (4.3) à (4.5).

THÉORÈME 4.1. *On suppose que les opérateurs* L, l, L^h, l^h, S^h, s^h *remplissent les conditions* A, B, C, E, F *et que* $f \in M_m(\Omega)$, $g \in N_m(D)$ *dans le problème* (2.2). *Les solutions* u_k^h, $k = 1, \ldots, m+1$, *du procédé itératif* (4.3) *à* (4.5) *admettent le développement*

$$u_k^h = u + \sum_{j=k}^{m} h^j v_{j,k} + \eta_k^h \quad \text{sur} \quad \bar{\Omega}_h. \tag{4.6}$$

Ici les fonctions $v_{j,k}$ *sont indépendantes de* h, $v_{j,k} \in P_{m-j}(\bar{\Omega})$, *et le reste* η_k^h *vérifie la majoration*

$$\|\eta_k^h\|_{\bar{\Omega}_h} \leqslant d_k\, h^{m+\beta}. \tag{4.7}$$

DÉMONSTRATION. On note que u_1^h coïncide avec la solution u^h du problème (2.3). D'après le théorème 2.1, on a donc pour elle le résultat voulu. On suppose la relation (4.6) et l'estimation (4.7) vraies pour un certain $k \geqslant 1$, et on se propose de les démontrer pour $k+1$.

Soit un ensemble quelconque de fonctions $v_{j,\,k+1} \in P_{m-j}(\bar{\Omega})$, $j = k+1, \ldots, m$, indépendantes de h. Avec ces fonctions et deux solutions u, u_{k+1}^h, on construit la fonction discrète

$$\eta_{k+1}^h = u_{k+1}^h - u - \sum_{j=k+1}^{m} h^j v_{j,\,k+1} \quad \text{sur} \quad \bar{\Omega}_h. \tag{4.8}$$

On exprime u^h à partir de cette relation et on la substitue dans le premier membre de (4.3). On porte dans le second membre le développement (4.6), il vient

$$L^h u + \sum_{j=k+1}^{m} h^j L^h v_{j,\,k+1} + L^h \eta_{k+1}^h =$$

$$= f + L^h u + \sum_{j=1}^{m} h^j L^h v_{j,\,k} + L^h \eta_k^h - S^h u - \sum_{j=1}^{m} h^j S^h v_{j,k} - S^h \eta_k^h.$$

On utilise la condition F et on effectue des simplifications:

$$\sum_{j=k+1}^{m} h^j v_{j,\,k+1} + L^h \eta_{k+1}^h =$$

$$= \sum_{j=k}^{m} h^j L^h v_{j,\,k} + L^h \eta_k^h - \sum_{j=k}^{m} h^j L\, v_{j,\,k} + \zeta_1^h - S^h \eta_k^h. \tag{4.9}$$

où

$$\| \zeta_1^h \|_{\breve{\Omega}_h} \leqslant c_{14} \, h^{m+\beta}. \tag{4.10}$$

On tient compte de la régularité de $v_{j,k}$, $v_{j,k+1}$ et on écrit par suite de la condition C:

$$L^h v_{j,k} = L v_{j,k} + \sum_{i=1}^{m-j} h^i A_{j,i} + \sigma_{j,k}^h \quad \text{sur} \quad \breve{\Omega}_h, \tag{4.11}$$

$$L^h v_{j,k+1} = L v_{j,k+1} + \sum_{j=1}^{m-j} h^i B_{j,i} + \sigma_{j,k+1}^h \quad \text{sur} \quad \breve{\Omega}_h. \tag{4.12}$$

Ici $A_{j,i}$, $B_{j,i} \in M_{m-j-i}(\Omega)$; $A_{j,i}$, $B_{j,i}$ ne dépendent pas de h, et les restes vérifient les inégalités

$$\| \sigma_{j,k}^h \|_{\breve{\Omega}_h} \leqslant c_{15} \, h^{m-j+\beta}, \qquad \| \sigma_{j,k+1}^h \|_{\breve{\Omega}_h} \leqslant c_{16} \, h^{m-j+\beta}. \tag{4.13}$$

Les développements (4.11), (4.12) aidant, on récrit (4.9):

$$\sum_{j=k+1}^{m} h^j \Big(L v_{j,k+1} + \sum_{i=1}^{j-k-1} B_{j-i,i} - \sum_{i=1}^{j-k} A_{j-i,i} \Big) + \\ + L^h \eta_{k+1}^h = L^h \eta_k^h - S^h \eta_k^h + \zeta_2^h \quad \text{sur} \quad \breve{\Omega}_h, \tag{4.14}$$

et (4.10), (4.13) impliquent l'estimation du reste

$$\| \zeta_2^h \|_{\breve{\Omega}_h} \leqslant c_{17} \, h^{m+\beta}. \tag{4.15}$$

On obtient de même à partir de (4.4)

$$\sum_{j=k+1}^{m} h^j \Big(l v_{j,k+1} + \sum_{i=1}^{j-k-1} b_{j-i,i} - \sum_{i=1}^{j-k} a_{j-i,i} \Big) + \\ + l^h \eta_{k+1}^h = l^h \eta_k^h - s^h \eta_k^h + \rho_2^h \quad \text{sur } D_h, \tag{4.16}$$

avec le reste évalué par

$$\| \rho_2^h \|_{D_h} \leqslant c_{18} \, h^{m+\beta}. \tag{4.17}$$

Les fonctions $a_{j,i}$, $b_{j,i}$ sont indépendantes de h; $a_{j,i}$, $b_{j,i} \in$ $\in N_{m-j-i}(D)$. Ces fonctions figurent dans les développements résultant de la condition C:

$$l^h v_{j,k} = l v_{j,k} + \sum_{i=1}^{m-j} h^i a_{j,i} + \varepsilon_{j,k}^h \quad \text{sur} \quad D_h, \tag{4.18}$$

$$l^h v_{j,k+1} = l v_{j,k+1} + \sum_{i=1}^{m-j} h^i b_{j,i} + \varepsilon_{j,k+1}^h \quad \text{sur} \quad D_h. \tag{4.19}$$

Ainsi, on vient d'obtenir, pour un jeu quelconque de $v_{j,\,k+1} \in P_{m-j}\,(\bar\Omega)$ et η^k définie par (4.8), les égalités (4.14), (4.16) avec les restes ζ_2^k, ρ_2^k évalués par (4.15), (4.17) respectivement.

On assimile $v_{j,\,k+1}$, $j = k+1, \ldots, m$, aux solutions des problèmes différentiels

$$\begin{aligned} Lv_{j,\,k+1} &= \sum_{i=1}^{j-k} A_{j-i,\,i} - \sum_{i=1}^{j-k-1} B_{j-i,\,i} \quad \text{dans} \quad \Omega, \\ lv_{j,\,k+1} &= \sum_{i=1}^{j-k} a_{j-i,\,i} - \sum_{i=1}^{j-k-1} b_{j-i,\,i} \quad \text{sur} \quad D. \end{aligned} \tag{4.20}$$

En particulier, le problème « déterminer la fonction $v_{k+1,\,k+1}$ » s'énonce comme suit :

$$\begin{aligned} Lv_{k+1,\,k+1} &= A_{k+1} \quad \text{dans} \quad \Omega, \\ lv_{k+1,\,k+1} &= a_{k,1} \quad \text{sur} \quad D. \end{aligned} \tag{4.21}$$

On a supposé que $v_{k,\,j}$ vérifient la condition C, d'où l'on a tiré les développements (4.11), (4.12). Aussi $A_{k,\,1}$, $a_{k,1}$ sont parfaitement définies par $v_{k,j}$, ne dépendent pas de h, et $A_{k,1} \in M_{m-k-1}\,(\Omega)$, $a_{k,\,1} \in N_{m-k-1}\,(D)$. Aux termes de la condition A, le problème (4.21) possède une solution unique $v_{k+1,\,k+1} \in P_{m-k-1}\,(\Omega)$.

Supposons qu'on connaît $v_{n,\,k+1} \in P_{m-n}\,(\bar\Omega)$ pour $n = k+1, \ldots$ $\ldots, j-1$, où $k+2 \leqslant j \leqslant m+1$. En vertu de la condition C, les développements (4.12), (4.19) ont lieu pour $n = k+1, \ldots, j-1$. Soit le problème (4.20) de déterminer la fonction $v_{j,\,k+1}$. Les termes $A_{j-i,\,i}$, $a_{j-i,\,i}$ sont définis, en raison de (4.11), par $v_{j,\,k}$ connues, et les termes $B_{j-i,\,i}$, $b_{j-i,\,i}$, $i = 1, \ldots, j-k-1$, sont parfaitement définis par les fonctions $v_{k+1,\,k+1}, \ldots, v_{j-1,\,k+1}$ en vertu de (4.12), $A_{j-i,\,i}$, $B_{j-i,\,i}$ étant dans $M_{m-j}\,(\Omega)$ et $a_{j-i,\,i}$, $b_{j-i,\,i}$ dans $N_{m-j}\,(D)$. La condition A implique donc que le problème (4.20) a une solution unique $v_{j,\,k+1} \in P_{m-j}\,(\bar\Omega)$ qui est évidemment indépendante de h.

Ainsi, on a défini les fonctions $v_{j,\,k+1}$, $j = k+1, \ldots, m$, ayant les propriétés voulues. L'ensemble de fonctions ainsi construites satisfait aux identités (4.14), (4.16). On simplifie ces identités moyennant les égalités (4.20) :

$$\begin{aligned} L^h\,\eta_{k+1}^h &= L^h\,\eta_k^h - S^h\,\eta_k^h + \zeta_2^h \quad \text{sur} \quad \Omega_h\,, \\ l^h\,\eta_{k+1}^h &= l^h\,\eta_k^h - s^h\,\eta_k^h + \rho_2^h \quad \text{sur} \quad D_h\,. \end{aligned}$$

Comment évaluer η_{k+1}^h ? Soient deux problèmes auxiliaires

$$\begin{aligned} L^h\,\varepsilon_1^h &= L^h\,\eta_k^h - S^h\,\eta_k^h \quad \text{sur} \quad \breve{\Omega}_h\,, \\ l^h\,\varepsilon_1^h &= l^h\,\eta_k^h - s^h\,\eta_k^h \quad \text{sur} \quad D_h\,; \end{aligned} \tag{4.22}$$

$$\begin{aligned} L^h\,\varepsilon_2^h &= \zeta_2^h \quad \text{sur} \quad \breve{\Omega}_h\,, \\ l^h\,\varepsilon_2^h &= \rho_2^h \quad \text{sur} \quad D_h\,. \end{aligned} \tag{4.23}$$

Dans les deux cas, il y a existence et unicité par suite de la condition B. On écrit η_{k+1}^h sous forme de somme $\varepsilon_1^h + \varepsilon_2^h$, d'où

$$\| \eta_{k+1}^h \|_{\bar{\Omega}_h} \leqslant \| \varepsilon_1^h \|_{\bar{\Omega}_h} + \| \varepsilon_2^h \|_{\bar{\Omega}_h}.$$

S'agissant de (4.22), la condition F entraîne l'estimation

$$\| \varepsilon_1^h \|_{\bar{\Omega}_h} \leqslant c_{12} \| \eta_k^h \|_{\bar{\Omega}_h} \leqslant c_{12}\, d_k\, h^{m+\beta}.$$

Quant à (4.23), on utilise (4.15), (4.17) et la condition B, il vient

$$\| \varepsilon_2^h \|_{\bar{\Omega}_h} \leqslant c\,(\| \zeta_2^h \|_{\breve{\Omega}_h} + \| \rho_2^h \|_{D_h}) \leqslant c\,(c_{17} + c_{18})\, h^{m+\beta}.$$

On réunit trois dernières inégalités et on établit pour η_{k+1}^h une estimation de la forme (4.7). Avec η_{k+1}^h de (4.8), on obtient le développement (4.6) pour $k + 1$. L'arbitraire laissé sur k démontre le théorème 4.1.

Ainsi, l'ordre de précision augmente avec chaque pas du procédé itératif (4.3) à (4.5):

$$\| u_k^h - u \|_{\bar{\Omega}_h} = O(h^k), \qquad k = 1, 2, \ldots, m,$$

$$\| u_{m+1}^h - u \|_{\bar{\Omega}_h} = O(h^{m+\beta}).$$

Quiconque voudra poursuivre les itérations dans le but d'obtenir une précision plus grande sera déçu dans son attente, car le procédé (4.3) à (4.5) donne un ordre de précision au plus égal à l'ordre d'approximation du schéma initial (4.1). Si l'on continue les itérations, deux situations sont en général à envisager. Si $c_{12} < 1$ dans la condition F, on montre l'existence et l'unicité pour le problème (4.1) et on évalue la stabilité par une inégalité de la forme (2.5). Dans ce cas, le procédé (4.3) à (4.5) converge avec la croissance de k vers la solution du problème (4.1), dont l'ordre de précision est celui de u_{m+1}^h. Comme le nombre d'opérations est proportionnel à k, l'efficacité de la méthode dans son ensemble diminue. Si c_{12} ne peut être inférieure à 1, le procédé itératif diverge avec k croissant, i.e. il donne lieu à une suite non bornée des u_k^h même si le problème

initial (4.1) est stable. Soit enfin $c_{12} = 1$. On obtient de plus une suite bornée des u_k^k d'ordre de précision au plus égal à l'ordre d'approximation du problème (4.1).

Ainsi, le fait de porter le nombre d'itérations au-delà d'une limite nécessaire n'améliore nullement la précision (et même il la réduit) tout en augmentant le coût des calculs. On conçoit que l'efficacité de la méthode s'en trouve amoindrie.

Voyons le cas fréquent où l'on a affaire, au lieu de la condition C, à la condition D dans laquelle les parties régulières des développements ne renferment que les puissances paires de h. On a le

THÉORÈME 4.2. *Si l'on est, pour les opérateurs* L, l, L^k, l^k, S^k, s^k, *dans les conditions* A, B, D, E, F *et* $f \in M_m(\Omega)$, $g \in N_m(D)$ *dans le problème* (2.2), *alors les solutions* u_k^k, $k = 1, \ldots, p+1$; $p = [m/2]$, *fournies par les itérations* (4.3) *à* (4.5) *admettent le développement*

$$u_k^k = u + \sum_{j=k}^{p} h^{2j} v_{j,k} + \eta_k^k \quad sur \quad \bar{\Omega}_k. \tag{4.24}$$

Ici $v_{j,k}$ *est indépendante de* h; $v_{j,k} \in P_{m-2j}(\bar{\Omega})$, *et le reste est évalué par*

$$\| \eta_k^k \|_{\bar{\Omega}_k} \leqslant d_k h^{m+\beta}. \tag{4.25}$$

La démonstration est analogue à celle du théorème 4.1 sauf que seules interviennent dans les calculs les puissances paires de h de la partie régulière. Avec les conditions du théorème 4.2, le procédé (4.3) à (4.5) garantit évidemment une croissance plus rapide de l'ordre de précision d'une itération à l'autre:

$$\| u_k^k - u \|_{\bar{\Omega}_k} = O(h^{2k}), \qquad k = 1, \ldots, p,$$

$$\| u_{p+1}^k - u \|_{\bar{\Omega}_k} = O(h^{m+\beta}).$$

Ainsi, l'ordre de précision donné par le théorème 4.1 est réalisé cette fois au bout des itérations environ deux fois moins nombreuses.

REMARQUE 1. Dans de nombreux problèmes, la condition aux limites $l^k u^k = g$ coincide sur D_k avec $s^k u^k = g$ et $lu = g$, auquel cas on remplace la condition de stabilité B par la condition B' du § 1.2 sans altérer pour autant les résultats des théorèmes 4.1 et 4.2. Les conditions E et F se simplifient de façon correspondante.

REMARQUE 2. S'agissant des problèmes à plusieurs variables, c'est la condition F qui s'avère la plus difficile à tester. La constant, c_{12} peut dépendre de h pour des normes simples d'emploi fréquente

Il en est alors également de d_k des théorèmes 4.1, 4.2, et $d_k = O((1 + c_{12}(h))^{k-1})$. Il en découle par suite du premier théorème

$$\| u_k^h - u \|_{\bar{\Omega}_h} = O(h^k (1 + c_{12}(h))^{k-1}), \qquad k = 1, \dots, m;$$

$$\| u_{m+1}^h - u \|_{\bar{\Omega}_h} = O(h^{m+\beta} (1 + c_{12}(h))^m).$$

Aussi le raffinement des itérations successives du schéma (4.3) à (4.5) n'apparaît, dans le cadre du théorème 4.1, qu'avec $c_{12}(h) = o(h^{-1})$. S'agissant du théorème 4.2, les estimations des pas d'itération s'écrivent

$$\| u_k^h - u \|_{\bar{\Omega}_h} = O(h^{2k} (1 + c_{12}(h))^{k-1}), \qquad k = 1, \dots, p;$$

$$\| u_{p+1}^h - u \|_{\bar{\Omega}_h} = O(h^{m+\beta} (1 + c_{12}(h))^p).$$

si bien que l'ordre de précision augmente dès que $c_{12}(h) = o(h^{-2})$.

1.5. Certains procédés d'extrapolation

L'extrapolation linéaire n'est pas le moyen unique pour accélérer la convergence des solutions approchées u^h pour $h \to 0$. On peut dire, de façon grossière, que son idée consiste à remplacer la fonction inconnue $u^h(x)$ (considérée comme fonction de la variable indépendante h) par le polynôme d'interpolation

$$f(h) = \sum_{i=0}^{k-1} \gamma_i h^{ip}, \qquad p = 1, 2, \tag{5.1}$$

dont la valeur $f(0)$ est prise pour limite approchée $\lim_{h \to 0} u^h(x)$. D'autres classes de fonctions d'interpolation conduisent naturellement à d'autres procédés d'extrapolation.

1.5.1. Extrapolation rationnelle

On prend pour fonctions d'interpolation les fonctions rationnelles de la forme

$$g(h) = \frac{\varphi(h^p)}{\psi(h^p)}, \qquad p = 1, 2. \tag{5.2}$$

le plus grand degré des polynômes $\varphi(t)$, $\psi(t)$ étant au plus $[k/2]$ et $[(k+1)/2]$ respectivement (la somme de ces nombres est égale à k). La classe des fonctions (5.2) contient en particulier les polynômes en h, et on s'en sert donc pour interpoler les développements de la forme (2.8) et (2.23).

On calcule évidemment les coefficients des polynômes φ, ψ par un procédé non linéaire assez laborieux. La valeur $g(0)$ est par contre obtenue par les récurrences simples de Stoer et Bulirsch (voir [76]). Soit le tableau d'extrapolation

$$\begin{array}{cccccc}
 & T_1^{(0)} & & & & \\
T_1^{(-1)} & & T_1^{(1)} \searrow & & & \\
 & T_2^{(0)} & \longrightarrow & T_1^{(2)} & & \\
T_2^{(-1)} & & T_2^{(1)} \nearrow & \cdot & T_1^{(k-2)} & T_1^{(k-1)} \\
\vdots & T_3^{(0)} & \vdots & \vdots & T_2^{(k-2)} & \\
\vdots & \vdots & \vdots & \vdots & & \\
T_{k-1}^{(-1)} & \vdots & T_{k-1}^{(1)} & T_{k-2}^{(2)} & & \\
 & T_k^{(0)} & & & &
\end{array}$$

On pose

$$T_j^{(-1)} = 0 \qquad \forall\, j = 2, \ldots, k;$$
$$T_j^{(0)} = u^{h_i} \qquad \forall\, j = 1, 2, \ldots, k,$$

et on calcule la suite récurrentielle

$$T_j^{(i)} = T_{j+1}^{(i-1)} + \frac{T_{j+1}^{(i-1)} - T_j^{(i-1)}}{\left(\dfrac{h_j}{h_{i+j}}\right)^p \left[1 - \dfrac{T_{j+1}^{(i-1)} - T_j^{(i-1)}}{T_{j+1}^{(i-1)} - T_{j+1}^{(i-2)}}\right] - 1}, \quad \begin{array}{l} i = 1, \ldots, k-j, \\ j = 1, \ldots, k-1. \end{array} \tag{5.3}$$

Le dernier nombre calculé $T_1^{(k-1)}$ du tableau donne la valeur $g(0)$ de la fonction rationnelle (5.2) qui parcourt les valeurs suivantes pour le jeu donné de h_i:

$$g(h_i) = u^k, \qquad i = 1, 2, \ldots, k. \tag{5.4}$$

On montre dans [76] que l'extrapolation rationnelle garantit en général à coût presque égal le même ordre de précision que l'extrapolation linéaire.

Comparons ces méthodes dans le cas de la fonction

$$u(h) = \alpha_0 + \alpha_1 h^2 + \alpha_2 h^4. \tag{5.5}$$

L'extrapolation linéaire sur deux valeurs h, $h/2$ du paramètre a pour résultat

$$u_L \equiv 4/3\ u(h/2) - 1/3\ u(h) = \alpha_0 - 1/4\ \alpha_2 h^4. \tag{5.6}$$

L'extrapolation rationnelle sur les mêmes valeurs de h aboutit à

$$u_R \equiv u(h/2) + \left[4\left(1 - \frac{u(h/2) - u(h)}{u(h)}\right) - 1\right]^{-1} [u(h/2) - u(h)] =$$
$$= \frac{3\,u(h)\,u(h/2)}{4\,u(h) - u(h/2)}. \tag{5.7}$$

On a, compte tenu de la forme explicite de u,

$$u_R = \alpha_0 + \frac{(\alpha_1^2 - \alpha_0\ \alpha_2)\ h^4/4 + 5\,\alpha_1\alpha_2\ h^6/16 + \alpha_2^2\ h^8/16}{\alpha_0 + 5\,\alpha_1\ h^2/4 + 21\,\alpha_2\ h^4/16}\,. \tag{5.8}$$

Soit $\alpha_0 \neq 0$. Considérons la partie principale de l'erreur d'extrapolation dans (5.6) et (5.8). La dernière relation se récrit

$$u_R = \alpha_0 + \left(\frac{\alpha_1^2}{4\alpha_0} - \frac{\alpha_2}{4}\right) h^4 + O\,(h^6)\,. \tag{5.9}$$

si bien que la partie principale de l'erreur commise dans l'extrapolation rationnelle vaut $\left(\frac{\alpha_1^2}{4\,\alpha_0} - \frac{\alpha_2}{4}\right) h^4$. Si α_0 et α_2 sont de même signe, elle est inférieure à l'erreur dans (5.6), et elle la dépasse dans le cas contraire. Comme la propriété de α_0 et α_2 d'être oui ou non de même signe est inconnue à priori, on ne saurait dire que l'une quelconque de ces méthodes est plus exacte que l'autre.

D'autre part, on voit intervenir dans les procédés d'extrapolation non linéaire des points singuliers tels que la précision de l'extrapolation baisse notablement dans leur voisinage. S'agissant de l'extrapolation rationnelle, cette perte de précision a lieu autour du point où la solution cherchée change de signe. Cette situation se présente pour la fonction-test (5.5) si l'on pose $\alpha_0 = 0$. Le résultat de l'extrapolation rationnelle est alors

$$u_R = \frac{1}{5}\,\alpha_1\ h^2 + O\,(h^4).$$

La valeur exacte est 0. Cela signifie que le procédé n'augmente pas l'ordre de précision devant $u(h)$ et $u(h/2)$. Quant à u_L, elle comporte l'erreur $O(h^4)$ quel que soit le signe de α_0.

1.5.2. Extrapolation exponentielle

On prend en qualité d'interpolants les combinaisons linéaires d'exponentielles

$$g\,(s) = a_0 + \sum_{i=1}^{k-1} a_i\, q_i^s\,, \tag{5.10}$$

a_i, q_i étant des paramètres indépendants. Il y a intérêt à calculer les limites (pour $s \to \infty$) des suites à partie principale de la forme (5.10) par la méthode (ou procédé δ^2) d'Aitken ou par la transformation de Shanks (voir [23]) qui en est une généralisation. Les deux techniques procèdent de la possibilité de calculer sans peine le coefficient a_0 dans $g(s)$ si les points d'interpolation sont équidistants.

On suppose $g(s)$ de (5.10) telle que

$$g(s) = h^{k_s}, \qquad s = n - k + 1,\ n - k + 2, \ldots, n + k - 1; \quad n \geqslant k. \tag{5.11}$$

On pose $\Delta g_i = g(i+1) - g(i)$ et on calcule deux déterminants d'ordre k:

$$D = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ \Delta g_{n-k+1} & \Delta g_{n-k+1} & \ldots & \Delta g_n \\ \Delta g_{n-k+2} & \Delta g_{n-k+3} & & \Delta g_{n+1} \\ \ldots & \ldots & \ldots & \ldots \\ \Delta g_{n-1} & \Delta g_n & \ldots & \Delta g_{n+k-2} \end{vmatrix},$$

$$D^* = \begin{vmatrix} g(n-k+1) & g(n-k+2) & \ldots & g(n) \\ \Delta g_{n-k+1} & \Delta g_{n-k+2} & \ldots & \Delta g_n \\ \Delta g_{n-k+2} & \Delta g_{n-k+3} & \ldots & \Delta g_{n+1} \\ \ldots & \ldots & \ldots & \ldots \\ \Delta g_{n-1} & \Delta g_n & \ldots & \Delta g_{n+k-2} \end{vmatrix}.$$

On définit a_0 par ces déterminants:

$$a_0 = D^*/D. \tag{5.12}$$

Si l'on effectue les calculs pour $n = k, k+1, k+2, \ldots$, on génère une nouvelle suite $a_0^{(k)}, a_0^{(k+1)}, \ldots$ dont on dit qu'elle est obtenue à partir de la suite u^{k_s} par le procédé de Shanks d'ordre $k - 1$. On note que si u^{k_s} possède pour $s \to \infty$ la partie principale de la forme (5.10), alors la vitesse de convergence de $a_0^{(n)}$, $n = k, k+1, \ldots$, est supérieure à celle de u^{k_s}.

Cette méthode donne pour $k = 1$ les formules d'Aitken connues

$$a_0 = \frac{\begin{vmatrix} g(n) & g(n+1) \\ \Delta g_{n-1} & \Delta g_n \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ \Delta g_{n-1} & \Delta g_n \end{vmatrix}} = \frac{g(n+1)\,g(n-1) - g^2(n)}{g(n+1) - 2g(n) + g(n-1)}. \tag{5.13}$$

Voyons si ces procédés sont opérants dans notre cas. On montre que les développements (2.8), (2.23) deviennent (5.10) par un choix des pas h_i. On pose notamment

$$h_s = h_0\, b^{-s}, \qquad s = 1, 2, \ldots \tag{5.14}$$

Ici h_0 est un pas initial et $b > 1$ le coefficient de resserrement des réseaux successifs. Pour que les réseaux possèdent le plus de points communs possible, on prend b égal à un entier. Le cas le plus simple de $b = 2$ a été discuté au § 1.3.

Avec h_s de (5.14), les développements (2.8), (2.23) se ramènent à la forme

$$u^{h_s} = u + \sum_{j=1}^{k-1} h_0^{jp} v_j (b^{-jp})^s + \eta^{h_s}, \tag{5.15}$$

avec $p = 1, 2$. On introduit les notations $a_0 = u$, $a_j = h_0^{jp} v_j$ et $q_j = b^{-jp}$ et on conclut que (5.10) approche (2.8), (2.23) avec une précision déterminée par la grandeur du reste η^{h_s}.

On fait la comparaison avec l'extrapolation linéaire et on utilise une fois de plus la fonction-test (5.5). Le procédé d'Aitken démarre avec trois valeurs $u(h_1)$, $u(h_2)$, $u(h_3)$. On pose conformément à (5.14) : $h_1 = h$, $h_2 = h/2$, $h_3 = h/4$, i.e. $b = 2$, $h_0 = h$. On a

$$u(h) = \alpha_0 + \alpha_1 h^2 + \alpha_2 h^4,$$
$$u(h/2) = \alpha_0 + \alpha_1 h^2/4 + \alpha_2 h^4/16,$$
$$u(h/4) = \alpha_0 + \alpha_1 h^2/16 + \alpha_2 h^4/256.$$

La valeur extrapolée s'écrit par suite de (5.13):

$$U_E = \alpha_0 + \frac{9\,\alpha_1\,\alpha_2\,h^4}{144\,\alpha_1 + 225\,\alpha_2 h^2} = \alpha_0 + \frac{1}{16}\alpha_2 h^4 + O(h^6). \tag{5.16}$$

Ainsi, la précision de $u(0)$ obtenue par le procédé d'Aitken est en effet $O(h^4)$. Mais si l'on extrapole linéairement sur deux valeurs $u(h/2)$, $u(h/4)$ de la fonction-test, on a pour résultat

$$u_L = \alpha_0 - \alpha_2 h^4/64.$$

Le coefficient de h^4 est 4 fois plus petit en valeur absolue que celui de la formule (5.16).

On voit donc que la méthode d'Aitken permet dans notre cas d'atteindre le même ordre de précision que l'extrapolation linéaire à condition d'utiliser plus de solutions approchées des problèmes auxiliaires. La cause en est les paramètres indépendants q_i de (5.10) dont la définition exige une information supplémentaire. Cela est encore plus vrai de la transformation de Shanks d'ordre $k > 1$. Il se peut en outre que le dénominateur de (5.12), (5.13) soit proche de 0, ce qui augmente sensiblement l'influence des erreurs de calcul.

1.5.3. ε-algorithme et ses généralisations

On recourt volontiers, pour accélérer la convergence des suites, à l' ε-algorithme (voir [140]). Ce dernier repose sur le tableau d'extrapolation

$$\begin{array}{ccccc}
 & \varepsilon_0^{(1)} & & & \\
\varepsilon_{-1}^{(2)} & & \varepsilon_1^{(1)} \searrow & & \\
 & \varepsilon_0^{(2)} & \longrightarrow & \varepsilon_2^{(1)} & \varepsilon_3^{(1)} \\
\varepsilon_{-1}^{(3)} & & \varepsilon_1^{(2)} \nearrow & & \vdots \\
 & \varepsilon_0^{(3)} & \vdots & \varepsilon_2^{(2)} & \vdots \\
\varepsilon_{-1}^{(4)} & & \vdots & \vdots & \vdots \\
\vdots & \varepsilon_0^{(4)} & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{array}$$

On se donne deux premières colonnes comme dans l'extrapolation rationnelle :

$$\varepsilon_{-1}^{(i)} = 0, \qquad i = 2, 3, \dots ; \qquad \varepsilon_0^{(i)} = u^{h_i}, \qquad i = 1, 2 \dots, \tag{5.17}$$

et les éléments des colonnes suivantes sont définis par l'égalité

$$\varepsilon_{s+1}^{(i)} = \varepsilon_{s-1}^{(i+1)} + \frac{1}{\varepsilon_s^{(i+1)} - \varepsilon_s^{(i)}}, \tag{5.18}$$

$$i = 1, 2, \dots; \quad s = 0, 1, \dots$$

Une étude approfondie de l'algorithme (voir [140]) a montré que les quantités $\varepsilon_{2k}^{(m)}$ donnent la transformation de Shanks d'ordre k, i.e. elles coïncident avec $a_0^{(m-k+1)}$ de l'algorithme (5.10) à (5.12). Si l'indice s est impair, alors $\varepsilon_s^{(m)}$ n'interviennent pas dans l'approximation de u^{h_s}.

On connaît plusieurs généralisations de l' ε-algorithme. On propose par exemple dans [73] le ρ-algorithme basé sur le même tableau d'extrapolation. Deux premières colonnes sont données par les formules (5.17) et les colonnes suivantes par

$$\varepsilon_{s+1}^{(i)} = \varepsilon_{s-1}^{(i+1)} + \frac{x_{s+i+1} - x_i}{\varepsilon_s^{(i+1)} - \varepsilon_s^{(i)}}, \qquad i = 1, 2, \dots ; \quad s = 0, 1, \dots ; \tag{5.19}$$

avec x_n une suite telle que $\lim\limits_{n \to \infty} x_n \to \infty$. Notre cas veut qu'on pose $x_n = (h_n)^{-p}$. On établit à l'aide de [73] que $\varepsilon_{2k}^{(1)}$ coïncide avec $T_1^{(2k-1)}$ de l'algorithme de Bulirsch-Stoer (5.3). Ainsi, le ρ-algorithme est une sorte d'extrapolation rationnelle économique.

On le trouve, en écriture un peu modifiée, dans [139] à côté d'un autre algorithme simple d'extrapolation rationnelle.

Ces algorithmes étant en fait des mises en œuvre économiques des extrapolations par des fonctions exponentielles et des fonctions rationnelles présentent sur l'extrapolation linéaire les mêmes avantages et offrent les mêmes inconvénients que ceux qu'on a étudiés plus haut.

1.6. Influence des erreurs de calcul

Si l'on traite sur ordinateur les problèmes aux différences

$$L^h u^h = f^h \quad \text{sur} \quad \bar{\Omega}_h , \qquad l^h u^h = g^h \quad \text{sur} \quad D_h , \tag{6.1}$$

on obtient de règle non la solution exacte u^h mais une fonction discrète $\tilde{u}^h$ affectée de l'erreur de calcul

$$\varepsilon^h = \tilde{u}^h - u^h \quad \text{sur} \quad \bar{\Omega}_h . \tag{6.2}$$

ε^h se compose des erreurs d'arrondi, des erreurs commises dans le calcul des coefficients et des seconds membres lorsqu'on remplace les opérations non arithmétiques par les opérations arithmétiques et des erreurs dues à la résolution approchée du système (6.1) par les méthodes itératives (s'agissant des systèmes non linéaires, on recourt, par exemple, aux itérations pour les problèmes aux limites discrets relatifs à des équations elliptiques). On amoindrit l'effet des erreurs de deux premiers types si l'on augmente le nombre de chiffres de la mantisse. On le fait dans des limites raisonnables au niveau du langage algorithmique, ce qui n'influe guère sur le temps de résolution du problème. S'agissant des erreurs du troisième type, il faut en outre augmenter le nombre d'itérations, et cet artifice risque de se répercuter de façon sensible sur le temps de calcul.

Voyons l'influence de l'erreur de calcul sur le résultat final de l'extrapolation linéaire. Soit $\tilde{u}^{h_k}$, $k = 1, 2, \ldots, s+1$, un jeu de solutions réelles du problème (6.1) avec paramètres de discrétisation $h_1, h_2, \ldots, h_{s+1}$. Les solutions sont affectées d'erreurs de calcul

$$\varepsilon^{h_k} = \tilde{u}^{h_k} - u^{h_k} .$$

Au lieu de la solution corrigée

$$U^H = \sum_{k=1}^{s+1} \gamma_k u^{h_k} , \tag{6.3}$$

on a en fait

$$\tilde{U}^H = \sum_{k=1}^{s+1} \gamma_k \tilde{u}^{h_k} = U^H + \sum_{k=1}^{s+1} \gamma_k \varepsilon^{h_k} . \tag{6.4}$$

Si l'on suppose que

$$h_k/h_{k+1} \geqslant 1 + d_1, \qquad k = 1, 2, \dots, s,$$

le lemme 2.3, § 7.2 entraîne

$$|\gamma_k| \leqslant d_2.$$

Aussi il résulte de (6.4)

$$|\tilde{U}^H - U^H| \leqslant d_2 \sum_{k=1}^{s+1} |\varepsilon^{h_k}|. \tag{6.5}$$

Si les paramètres h_k ne sont pas trop rapprochés deux à deux, l'erreur de calcul dans l'extrapolation linéaire est donc égale en ordre de grandeur à la somme des erreurs de calcul sur les solutions $\tilde{u}^{h_k}$ utilisées. Nos raisonnements ne tiennent pas compte des erreurs d'arrondi dans la somme (6.3). On établit sans peine leur influence négligeable.

Il en va autrement des erreurs de calcul dans la méthode des différences d'ordre supérieur. On considère, avec les notations du § 1.4, les $m + 1$ pas du procédé

$$\begin{aligned} L^h u_{k+1}^h &= f + L^h u_k^h - S^h u_k^h \quad \text{sur} \quad \check{\Omega}_h, \\ l^h u_{k+1}^h &= g + l^h u_k^h - s^h u_k^h \quad \text{sur} \quad D_h, \end{aligned} \tag{6.6}$$

qui vérifie la condition F. On prend pour approximation initiale $\tilde{u}_0^h = 0$ sur $\bar{\Omega}_h$.

On a déjà dit que le problème (6.6) est résolu pour chaque $k = 0, 1, \dots, m$ de façon approchée, si bien qu'on utilise non u^{h_k} mais $\tilde{u}_k^h$, l'erreur de calcul étant $\eta_k = \tilde{u}_k^h - u_k^h$. Le problème (6.6) fait donc place à

$$\begin{aligned} L^h v_{k+1} &= f + L^h u_k^h - S^h \tilde{u}_k^h \quad \text{sur} \quad \check{\Omega}_h, \\ l^h v_{k+1} &= g + l^h \tilde{u}_k^h - s^h \tilde{u}_k^h \quad \text{sur} \quad D_h \end{aligned} \tag{6.7}$$

dont la solution diffère de u_{k+1}^h par la quantité $\rho_{k+1} = v_{k+1} - u_{k+1}^h$. On a pour ce problème

$$\begin{aligned} L^h \rho_{k+1} &= L^h \eta_k - S^h \eta_k \quad \text{sur} \quad \check{\Omega}_h, \\ l^h \rho_{k+1} &= l^h \eta_k - s^h \eta_k \quad \text{sur} \quad D_h. \end{aligned}$$

La condition F entraîne

$$\| \rho_{k+1} \|_{\bar{\Omega}_h} \leqslant c_{12} \| \eta_k \|_{\bar{\Omega}_h}.$$

Mais le système (6.7) est résolu approximativement lui aussi. On a donc finalement $\tilde{u}_{k+1}$ (au lieu de v_{k+1}) avec l'erreur de calcul $\varepsilon_{k+1} =$

$= \tilde{u}_{k+1} - v_{k+1}$. Ainsi, le k-ième pas fournit $\tilde{u}^h_{k+1}$ avec l'erreur de calcul

$$\eta_{k+1} = \tilde{u}^h_{k+1} - u^h_{k+1}$$

qui admet la majoration

$$\| \eta_{k+1} \|_{\bar{\Omega}_h} \leqslant c_{12} \| \eta_k \|_{\bar{\Omega}_h} + \| \varepsilon_{k+1} \|_{\bar{\Omega}_h}. \tag{6.8}$$

On fait l'hypothèse que le problème (6.7) est résolu pour tout $k = 0, 1, \ldots, m$ avec une erreur de calcul au plus égale en norme au nombre $\delta > 0$: $\| \varepsilon_{k+1} \|_{\bar{\Omega}_h} \leqslant \delta$, $k = 0, 1, \ldots, m$, et on demande l'erreur totale. Dans le cas $k = 0$, on prend $\tilde{u}^h_0 = u^h_0 = 0$, donc $\| \eta_0 \|_{\bar{\Omega}_h} = 0$. Aussi l'estimation (6.8) entraîne

$$\| \eta_{m+1} \|_{\bar{\Omega}_h} \leqslant \delta \frac{c_{12}^{m+1} - 1}{c_{12} - 1} \quad \text{pour} \quad c_{12} \neq 1,$$

$$\| \eta_{m+1} \|_{\bar{\Omega}_h} \leqslant \delta (m + 1) \quad \text{pour} \quad c_{12} = 1.$$

La constante c_{12} est en général plusieurs fois plus grande que 1, auquel cas on résout le problème (6.7) pour chaque $k = 0, 1, \ldots m$ avec une réserve de précision. Une tactique commode est de le résoudre avec altération de la précision de sorte que

$$\| \varepsilon_{k+1} \|_{\bar{\Omega}_h} \leqslant \delta / c_{12}^{m-k+1}, \qquad k = 0, 1, \ldots, m.$$

L'estimation de l'erreur de calcul sur la solution définitive est alors indépendante de c_{12}:

$$\| \eta_{m+1} \|_{\bar{\Omega}_h} \leqslant \delta (m + 1).$$

Ainsi, la méthode des différences d'ordre supérieur est plus sensible aux erreurs de calcul que l'extrapolation linéaire de Richardson.

CHAPITRE 2

ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES DU PREMIER ORDRE

Lorsqu'on résout numériquement des équations différentielles ordinaires, on demande de règle des solutions précises au possible par rapport au pas du réseau sur lequel le problème différentiel est réduit au problème aux différences. On connaît de nombreuses méthodes qui permettent d'approcher la solution régulière du problème différentiel avec une précision donnée. La méthode de Runge-Kutta occupe parmi celles-ci une place spéciale tant du point de vue de son champ d'applications qu'en ce qui concerne la possibilité qu'elle offre de traiter de façon uniforme et particulièrement simple la mise en œuvre de l'algorithme de résolution des problèmes. Ajoutez-y une théorie bien élaborée, et vous comprendrez pourquoi cette méthode fait l'objet d'une grande attention en Analyse numérique.

Depuis plusieurs années, les mathématiciens s'intéressent aux algorithmes implicites et semi-implicites pour des systèmes d'équations différentielles ordinaires, et notamment pour les cas où les dérivées d'ordre supérieur sont munies d'un paramètre petit. De tels systèmes sont dits « rigides », et ils sont d'ordinaire caractérisés par la propriété de la solution de croître (ou décroître) vite. de façon exponentielle, au début de l'intervalle de variation de la variable indépendante pour ne varier ensuite que d'une manière assez régulière. Le fait d'utiliser un algorithme uniforme adaptatif est à l'origine de plusieurs procédures de calcul de valeur dont les méthodes de Rosenbrock qui sont dignes de mention (on les interprète comme régularisation originale des algorithmes de Runge-Kutta connus; voir [14], [126]).

Dans ce chapitre, nous utiliserons des algorithmes très simples. Fidèles à l'idée de l'extrapolation de Richardson, nous les exploiterons pour plusieurs réseaux différents. Dans de nombreux cas, la combinaison linéaire des solutions obtenues définies sur le réseau initial constitue une solution aussi exacte que sa régularité le permet. L'exactitude limite est déterminée par le schéma de mise en œuvre et par l'analyse de la précision qui sont basés sur le développement de la solution suivant les puissances d'un paramètre petit

(ce paramètre est en fait le pas du réseau). Aussi la dérivabilité de la solution détermine l'approximation numérique avec une précision donnée.

Ce sont les équations différentielles ordinaires du premier ordre qui sont le mieux étudiées du point de vue de l'extrapolation. On trouve dans [94] la liste des ouvrages théoriques antérieurs à 1971. Les auteurs [55], [82], [129], [130] font une étude comparée sérieuse des mises en œuvre numériques de nombreuses méthodes. Il se trouve que contrairement aux autres procédés pour des systèmes non rigides, diverses modifications de l'extrapolation sont le plus efficaces si l'on veut une précision très élevée ou si le pas d'intégration est susceptible d'une estimation simple qui garantit la précision donnée. Si cette estimation fait défaut, la plus grande partie du temps de résolution va à trouver le pas d'intégration initial. Quant à la méthode d'extrapolation même, elle se distingue par une grande économie.

2.1. Schéma de Crank-Nicholson

Dans le chapitre précédent, nous avons eu affaire à une équation linéaire des plus simples. Dans la suite, on étudiera une équation différentielle générale du premier ordre. On suppose que les données du problème satisfont à certaines conditions qui garantissent une solution suffisamment régulière et, partant, des solutions approchées d'ordre de précision élevé.

Fin du paragraphe, on examinera deux questions spéciales à l'extrapolation de Richardson. Ce sont, *primo*, le rôle des polynômes d'interpolation de Lagrange dans la construction d'une solution très précise associée aux points qui ne sont pas des nœuds, et, *secundo*, le rapport de pas des réseaux successifs.

2.1.1. Extrapolation pour des problèmes non rigides

Soit le problème

$$\frac{du}{dt} = f(t, u), \qquad t \in (0, 1), \tag{1.1}$$

$$u(0) = u_0. \tag{1.2}$$

On introduit les classes $C^m(\bar{\Omega})$ des fonctions m fois continûment dérivables sur l'ensemble $\bar{\Omega}$. On munit les éléments de $C^m[0, 1]$ d'une norme définie par

$$\|\varphi\|_{C^m[0,1]} = \max_{0 \leqslant k \leqslant m} \max_{0 \leqslant t \leqslant 1} |\varphi^{(k)}(t)|.$$

Le second membre de (1.1) est supposé être

$$f \in C^r([0, 1] \times (-\infty, \infty)), \tag{1.3}$$

$r \geqslant 2$ étant une constante prenant des valeurs entières et

$$\sup_{\substack{t \in [0, 1] \\ y \in (-\infty, \infty)}} \left| \frac{\partial^i f}{\partial y^i}(t, y) \right| \leqslant c_1, \qquad i = 0, 1, 2. \tag{1.4}$$

La condition (1.4) n'étant pas remplie par des fonctions très simples telles que par exemple $f(t, y) = y^2$ paraît trop sévère. On la vérifie par suite de (1.3) si le problème proposé admet une solution bornée. En effet, on fait l'hypothèse que la solution u de (1.1) à (1.3) est majorée par

$$\| u \|_{C[0, 1]} \leqslant c_2, \tag{1.5}$$

auquel cas on définit $f(t, y)$ à l'extérieur du rectangle $[0, 1] \times [-c_2, c_2]$ de façon qu'elle soit toujours de classe $C^r([0, 1] \times (-\infty, \infty))$ et qu'on reste dans la condition (1.4) (à cet effet on pose par exemple $f(t, y) = 0 \ \forall |y| \geqslant 2c_2$ et $t \in [0, 1]$). Il est clair que le problème (1.1) à (1.3) avec f ainsi modifié possède une solution u (qui est d'ailleurs unique).

La réciproque est également vraie (voir par exemple [77] : les conditions (1.3), (1.4) impliquent l'existence et l'unicité pour le problème (1.1), (1.2), et la continuité de la solution u sur $[0, 1]$ entraîne

$$u \in C^{r+1}[0, 1]. \tag{1.6}$$

On construit le réseau régulier

$$\omega_\tau = \{t_j = j\tau, \ j = 0, 1, \ldots, M\} \tag{1.7}$$

de pas $\tau = 1/M$ et on introduit les points médians des intervalles partiels

$$\bar{\omega}_\tau = \{t_{j+1/2} = (j + 1/2)\tau, \quad j = 0, 1, \ldots, M-1\}. \tag{1.8}$$

La résolution numérique du problème (1.1), (1.2) sera effectuée avec le schéma de Crank-Nicholson (connu également sous le nom de méthode des rectangles implicite)

$$u_{\bar{t}}^\tau = f(t, u_{\bar{t}}^{\bar{\tau}}) \quad \text{sur} \quad \bar{\omega}_\tau, \tag{1.9}$$

$$u^\tau(0) = u_0. \tag{1.10}$$

On démontre que la condition (1.4) garantit, pour τ suffisamment petits , la compatibilité du système obtenu d'équations non linéaires. On propose un procédé de calcul des valeurs successives de

$u^\tau(t)$ pour $t = \tau, 2\tau, 3\tau, \ldots$ Supposons qu'on connaît $u^\tau(t)$ pour un certain $t \in \omega_\tau$. On écrit l'équation (1.9) associée à $t + \tau/2$:

$$\frac{u^\tau(t+\tau) - u^\tau(t)}{\tau} = f\left(t + \frac{\tau}{2}, \frac{u^\tau(t+\tau) + u^\tau(t)}{2}\right). \quad (1.11)$$

On la résout par rapport à $u^\tau(t+\tau)$ par la méthode de Newton [96] pour les équations non linéaires de la forme

$$p(x) = 0, \quad (1.12)$$

avec p une fonction régulière d'une variable réelle. Soit x_0 l'approximation initiale. Les valeurs suivantes sont calculées par la formule

$$x_{n+1} = x_n - [p'(x_n)]^{-1} p(x_n), \qquad n = 0, 1, \ldots \quad (1.13)$$

La suite x_n converge sous certaines conditions vers la solution du problème (1.12). Voici des conditions suffisantes de convergence (voir [23]).

THÉORÈME 1.1. *On suppose que*

1) *la fonction $p(x)$ est définie dans le disque fermé*

$$|x - x_0| \leqslant \delta, \quad (1.14)$$

est deux fois dérivable dans le disque, la dérivée seconde étant bornée:

$$|p''(x)| \leqslant K, \qquad |x - x_0| \leqslant \delta; \quad (1.15)$$

2) $$|p'(x_0)| \geqslant B; \quad (1.16)$$

3) $$|p(x_0)/p'(x_0)| \leqslant \eta; \quad (1.17)$$

4) *les quantités K, B, η vérifient la condition*

$$h = \frac{K\eta}{B} \leqslant 1/2; \quad (1.18)$$

5) *on a pour δ*

$$\frac{1 - \sqrt{1-2h}}{h}\eta \leqslant \delta. \quad (1.19)$$

Alors

1) *l'équation $p(x) = 0$ possède une solution unique x^* dans le disque* (1.14);

2) *dans la méthode* (1.13), *l'approximation x_n est générée pour tout n, x_n sont toutes dans le disque* (1.14) *et $x_n \to x^*$ avec $n \to \infty$.*

Voyons si les conditions du théorème sont justes pour l'équation (1.11). On récrit (1.11):

$$p(x) \equiv \frac{x - x_0}{\tau} - f\left(t + \frac{\tau}{2}, \frac{x + x_0}{2}\right) = 0, \qquad x_0 = u^\tau(t). \quad (1.20)$$

La fonction $p(x)$ est définie sur la droite réelle tout entière, et la dérivée seconde est évaluée pour tout $x \in (-\infty, \infty)$ par

$$|p''(x)| \leqslant \frac{1}{4}\left|\frac{\partial^2 f}{\partial u^2}\left(t+\frac{\tau}{2}, \frac{x+x_0}{2}\right)\right| \leqslant \frac{c_1}{4}.$$

Ainsi , on choisit la constante δ aussi grande qu'on le veut et $K = c_1/4$. Pour qu'on soit dans la deuxième condition, il suffit de prendre $\tau < 2/c_1$. On pose par exemple $\tau \leqslant 1/c_1$, auquel cas

$$|p'(x_0)| = \left|\frac{1}{\tau} - \frac{1}{2}\frac{\partial f}{\partial u}\left(t+\frac{\tau}{2}, x_0\right)\right| \leqslant c_1/2.$$

On choisit donc la constante $B = c_1/2$. Il reste à trouver η. On considère l'inégalité

$$p(x_0)/p'(x_0)| = \frac{\tau\,|f(t+\tau/2, x_0)|}{1-\frac{\tau}{2}\frac{\partial f}{\partial u}\left(t+\frac{\tau}{2}, x_0\right)} \leqslant$$

$$\leqslant 2\tau\left|f\left(t+\frac{\tau}{2}, x_0\right)\right| \leqslant 2\tau c_1,$$

si bien qu'on pose $\eta = 2\tau c_1$. On a (1.18) sous la condition nécessaire

$$h = \frac{c_1}{4}\cdot 2\tau c_1 \cdot \frac{2}{c_1} = \tau c_1 \leqslant \frac{1}{2}.$$

D'où $\tau \leqslant 1/2c_1$. L'inégalité (1.19) est nécessairement vérifiée par suite de l'arbitraire laissé sur δ. Il suffit donc, pour être dans toutes les hypothèses du théorème, de prendre

$$\tau \leqslant 1/2\, c_1. \tag{1.21}$$

Dans ce cas, le théorème entraîne que l'équation (1.20) et, partant, (1.11) possèdent une solution unique x^* qu'on désigne par $u^\tau(t+\tau)$.

La méthode de Newton (1.13), (1.20) est à convergence quadratique [23]. Etant donnée l'approximation initiale $x_0 = u^\tau(t)$, on a l'estimation $|x^* - x_0| = O(\tau)$, et la convergence quadratique implique

$$|x^* - x_1| = O(\tau^2),$$
$$|x^* - x_2| = O(\tau^4),$$
$$|x^* - x_3| = O(\tau^8) \ldots$$

Si l'on est dans la condition (1.21) et si les solutions sont régulières, la méthode de Newton aboutit donc au bout de pas peu nombreux. On doit cependant choisir un critère d'arrêt. On convient par exem-

ple de stopper les itérations après qu'on a pour $\varepsilon > 0$ donné suffisamment petit

$$|p(x_n)| \leqslant \varepsilon, \tag{1.22}$$

puis on prend x_n pour valeur approchée $u^\tau(t+\tau)$. Dans ce cas, le problème (1.9), (1.10) est en fait remplacé par le problème proche

$$v_{\bar{t}}^\tau = f(t, v_{\hat{t}}^\tau) + \rho \quad \text{sur} \quad \bar{\omega}_\tau \tag{1.23}$$

$$v^\tau(0) = u_0, \tag{1.24}$$

où

$$|\rho(t)| \leqslant 2\varepsilon \quad \text{sur} \quad \bar{\omega}_\tau, \tag{1.25}$$

ρ comprenant les erreurs d'arrondi, le résidu de la méthode de Newton et les erreurs commises en calculant f. On démontre que la solution v^τ de ce problème diffère peu de la solution u^τ du problème aux différences initial.

THÉORÈME 1.2. *On suppose que la fonction f présente la propriété* (1.4) *et que τ vérifie l'estimation* (1.21). *On a*

$$\| u^\tau - v^\tau \|_{C,\tau} \leqslant \frac{\varepsilon}{c_1}(e^{2c_1} - 1). \tag{1.26}$$

DÉMONSTRATION. On montre la justesse de l'inégalité

$$|u^\tau(t) - v^\tau(t)| \leqslant \frac{\tau\varepsilon}{1-\tau c_1/2} \sum_{j=0}^{t/\tau - 1} \left(\frac{1+\tau c_1/2}{1-\tau c_1/2}\right)^j. \tag{1.27}$$

Elle est évidente pour $t = 0$ car $v^\tau(0) - u^\tau(0) = 0$ et la somme du second membre est nulle (on en a convenu au § 1.2). Supposons-la vraie pour un $t \in \omega_\tau$ et démontrons-la pour $t+\tau$. On trouve la différence entre (1.23) et (1.9) et on prend le module des deux membres, il vient

$$|v_{\bar{t}}^\tau - u_{\bar{t}}^\tau| \leqslant |f(t, v_{\hat{t}}^\tau) - f(t, u_{\hat{t}}^\tau)| + |\rho|. \tag{1.28}$$

La condition (1.4) implique l'inégalité simple

$$|f(t, v_{\hat{t}}^\tau) - f(t, u_{\hat{t}}^\tau)| = |v_{\hat{t}}^\tau - u_{\hat{t}}^\tau| \left| \frac{\partial f}{\partial u}(t, \zeta) \right| \leqslant c_1 |v_{\hat{t}}^\tau - u_{\hat{t}}^\tau|.$$

Avec cette inégalité et l'estimation (1.25), l'inégalité (1.28) devient

$$|v_{\bar{t}}^\tau - u_{\bar{t}}^\tau| \leqslant c_1 |v_{\hat{t}}^\tau - u_{\hat{t}}^\tau| + \varepsilon.$$

On associe cette relation au point $t + \tau/2$, il vient après plusieurs transformations élémentaires

$$(1 - \tau c_1/2)\,|v^\tau(t+\tau) - u^\tau(t+\tau)| \leqslant |1 + \tau c_1/2)\,(v^\tau(t) - u^\tau(t)\,| + \tau\varepsilon.$$

Le facteur $1 - \tau c_1/2$ est positif par suite de (1.21). On divise membre à membre par ce facteur et on utilise (1.27):

$$|v^\tau(t+\tau) - u^\tau(t+\tau)| \leqslant \frac{1+\tau c_1/2}{1-\tau c_1/2}\,|v^\tau(t) - u^\tau(t)| + \frac{\tau\varepsilon}{1+\tau c_1/2} \leqslant$$
$$\leqslant \frac{\tau\varepsilon}{1-\tau c_1/2}\sum_{j=0}^{t/\tau}\left(\frac{1+\tau c_1/2}{1-\tau c_1/2}\right)^j .$$

Ainsi, l'inégalité (1.27) a bien lieu. On simplifie son second membre moyennant l'inégalité

$$\frac{1+x}{1-x} \leqslant e^{4x}, \qquad 0 \leqslant x \leqslant 1/4.$$

et la propriété $\tau c_1/2 \leqslant 1/4$ résultant de (1.21):

$$|v^\tau(t) - u^\tau(t)| \leqslant \frac{\tau\varepsilon}{1-\tau c_1/2}\left[\left(\frac{1+\tau c_1/2}{1-\tau c_1/2}\right)^{t/\tau} - 1\right]\Bigg/\left[\frac{1+\tau c_1/2}{1-\tau c_1/2} - 1\right] \leqslant$$
$$\leqslant \frac{\varepsilon}{c_1}\,(e^{2tc_1} - 1).$$

On maximise par rapport à $t \in \bar\omega_\tau$ et on obtient l'estimation (1.26), c.q.f.d.

Quitte à compliquer la démonstration, on diminue de moitié l'exposant de la fonction exponentielle de (1.26).

Soit le problème aux différences (1.9), (1.10). On démontre que sa solution est développable en série procédant suivant les puissances paires de τ (voir également [108]).

THÉORÈME 1.3. *On suppose que le problème* (1.1), (1.2) *remplit les conditions* (1.3), (1.4). *La solution du problème aux différences* (1.9), (1.10) *admet sous la condition* (1.21) *le développement*

$$u^\tau = u + \sum_{j=1}^{m} \tau^{2j} v_j + \eta^\tau \quad sur \quad \bar\omega_\tau, \tag{1.29}$$

avec $m = [(r-1)/2]$, v_j *indépendantes de* τ, $v_j \in C^{r-2j+1}[0,1]$, *et le reste évalué par*

$$\|\eta^\tau\|_{C,\tau} \leqslant c_3\tau^r. \tag{1.30}$$

DÉMONSTRATION. On pose $v_0 \equiv u$ et on cherche m fonctions v_j à partir des problèmes différentiels

$$v_j' - \frac{\partial f}{\partial y}(t,u)\,v_j = \sum_{k=1}^{j}\left(-\frac{v_{j-k}^{(2k+1)}}{4^k(2k+1)!} + \frac{v_{j-k}^{(2k)}}{4^k(2k)!}\,\frac{\partial f}{\partial y}(t,u) + \right.$$
$$\left. + \sum_{l=2}^{j}\frac{1}{l!}\,\frac{\partial^l f}{\partial y^l}(t,u)\sum_{i_1+\ldots+i_l=j}\ \prod_{i=1}^{l}\left(\sum_{k=0}^{i_i}\frac{v_{i_i-k}^{(2k)}}{4^k(2k)!}\right)\right) \text{ sur } (0,1), \tag{1.31}$$

$$v_j(0) = 0, \qquad j = 1, 2, \ldots, m. \tag{1.32}$$

Le problème correspondant à v_1 s'écrit par exemple

$$v_1' - \frac{\partial f}{\partial y}(t, u)\, v_1 = -u'''/24 + \frac{\partial f}{\partial y}(t, u)\, u''/8, \tag{1.33}$$

$$v_1(0) = 0. \tag{1.34}$$

Le second membre de l'équation contient uniquement la fonction $v_0 \equiv u$ et est $r - 2$ fois continûment dérivable. Le problème étant linéaire admet une solution unique $v_1 \in C^{r-1}[0, 1]$. On suppose définies les fonctions $v_1, \ldots, v_{j-1}$, et $v_k \in C^{r-2j+1}[0, 1]$. Le second membre de l'équation (1.31) associée à l'indice j renferme v_j d'indice au plus égal à $j - 1$ et est $r - 2j$ fois continûment dérivable par rapport à $t \in [0, 1]$. Aussi le problème linéaire (1.31), (1.32) pour v_j a une solution unique de classe $C^{r-2j+1}[0, 1]$. On a donc trouvé toutes les fonctions v_j.

On introduit la fonction

$$w = \sum_{j=0}^{m} \tau^{2j} v_j \quad \text{sur} \quad \bar\omega_\tau$$

et on la porte dans l'opérateur aux différences de (1.9):

$$w_{\bar t} - f(t, w_{\bar t}) = \sum_{j=0}^{m} \tau^{2j} (v_j)_{\bar t} - f\Big(t, \sum_{j=0}^{m} \tau^{2j} (v_j)_{\bar t}\Big). \tag{1.35}$$

On utilise les développements du lemme 1.1, § 7.1 pour transformer les termes du second membre:

$$\sum_{j=0}^{m} \tau^{2j} (v_j)_{\bar t} = \sum_{j=0}^{m} \tau^{2j} \sum_{k=0}^{m} \tau^{2k} \frac{v_j^{(2k+1)}}{4^k (2k+1)!} + h^r \theta_1 = \sum_{j=0}^{m} \tau^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k+1)}}{4^k (2k+1)!} + h^r \theta_1,$$

$$\sum_{j=0}^{m} \tau^{2j} (v_j)_{\bar t} = \sum_{j=0}^{m} \tau^{2j} \sum_{k=0}^{j} \frac{v_{j-k}}{4^k (2k)!} + h^r \theta_2, \tag{1.36}$$

où $|\theta_i| \leqslant c_4$ pour $i = 1, 2$ sur $\bar\omega_\tau$. On développe $f(t, w_{\bar t})$ en formule de Taylor:

$$f(t, w_{\bar t}) = f\Big(t, \sum_{j=0}^{m} \tau^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k)}}{4^k (2k)!} + h^r \theta_2\Big)\frac{\partial f}{\partial y}(t, \zeta) = f(t, u) + \sum_{l=1}^{m} \Big\{ \sum_{j=1}^{m} \tau^{2j} \times \sum_{k=0}^{j} \frac{v_{j-k}^{(2k)}}{4^k (2k)!} \Big\}^l \frac{1}{l!} \frac{\partial^l f}{\partial y^l}(t, u) + h^r \theta_3.$$

La propriété de borne des fonctions v_j, f et de leurs dérivées d'ordre correspondant entraîne $|\theta_3| \leqslant c_5$ sur $\bar{\omega}_\tau$. On élève à la puissance les accolades et on réduit les termes contenant les mêmes puissances de τ:

$$f(t, w_\tau) = f(t, u) + \sum_{j=1}^{m^2} \tau^{2j} \left\{ \sum_{l=1}^{m} \frac{1}{l!} \times \right.$$

$$\left. \times \frac{\partial^l f}{\partial y^l}(t, u) \sum_{t_1+\ldots+t_l=j} \prod_{i=1}^{l} \left(\sum_{k=0}^{t_i} \frac{v_{t_i-k}^{(2k)}}{4^k (2k)!} \right) \right\} + h^r \theta_3 .$$

On conserve les termes en τ^{2j}, $j \leqslant m$, et on écrit les termes restants comme $h^r \theta_4$, où $|\theta_4| \leqslant c_6$ sur $\bar{\omega}_\tau$. L'égalité obtenue et les relations (1.35), (1.36) donnent

$$w_\tau - f(t, w_\tau) = u' - f(t, u) + \sum_{j=1}^{m} \tau^{2j} \left\{ \sum_{k=0}^{j} \frac{v_{j-k}^{(2k+1)}}{4^k (2k+1)!} - \right.$$

$$\left. - \sum_{l=1}^{j} \frac{1}{l!} \frac{\partial^l f}{\partial y^l}(t, u) \sum_{t_1+\ldots+t_l=j} \prod_{i=1}^{l} \left(\sum_{k=0}^{t_i} \frac{v_{t_i-k}^{(2k)}}{4^k (2k)!} \right) \right\} +$$

$$+ h^r \theta_1 - h^r \theta_3 + h^r \theta_4 .$$

Tous les termes dans les accolades se réduisent par suite de la définition des fonctions v_j, et l'équation (1.1) l'entraîne en ce qui concerne $u' - f(t, u)$, il vient finalement

$$w_\tau = f(t, w_\tau) + h^r \theta_1 - h^r \theta_3 + h^r \theta_4 \text{ sur } \bar{\omega}_\tau . \qquad (1.37)$$

Les conditions initiales (1.2), (1.10), (1.32) impliquent de plus $w(0) = u^h(0) = u(0)$. La fonction w_τ obéit donc au théorème 1.2, d'où l'estimation $\|\eta^\tau\|_{C,\tau} \leqslant c_3 \tau^r$ pour $\eta^\tau = u^h - w$, i.e. on a la dernière affirmation du théorème.

On utilise le développement obtenu pour augmenter la précision par l'extrapolation de Richardson décrite au § 1.3. On suppose réalisées les conditions (1.3), (1.4) et on pose $s = [(r-1)/2]$. On construit pour $0 < N_1 < \ldots < N_{s+1}$ entiers fixés les réseaux $\bar{\omega}_{\tau_k}$ de pas $\tau_k = 1/(N_k M)$, avec M un entier naturel qui peut être en principe aussi grand qu'on le veut et tel que

$$M \geqslant 2\, c_1/N_1 . \qquad (1.38)$$

Cette condition garantit la possibilité du problème aux différences (1.9), (1.10) sur chaque réseau $\bar{\omega}_{\tau_k}$. Toutes les solutions u^{τ_k} sont définies sur le réseau de pas $\tau = 1/M$.

Soit le système

$$\sum_{k=1}^{s+1} \gamma_k = 1, \qquad \sum_{k=1}^{s+1} \gamma_k \tau_k^{2j} = 0, \qquad j = 1, \ldots, s. \tag{1.39}$$

Comme τ_k sont distincts deux à deux, le système est non dégénéré et admet une solution unique. On prend la combinaison linéaire sur $\bar{\omega}_\tau$

$$U^H = \sum_{k=1}^{s+1} \gamma_k u^{\tau_k} \tag{1.40}$$

et on montre que l'ordre de précision de U^H est pour $\tau \to 0$ supérieur à celui de chaque u^{τ_k}.

THÉORÈME 1.4. *On suppose qu'on est, pour le problème* (1.1), (1.2), *dans les conditions* (1.3), (1.4). *La solution corrigée* (1.40) *avec les poids obtenus à partir du système* (1.39) *vérifie sous la condition* (1.38) *l'estimation*

$$\| U^H - u \|_{C,\tau} \leqslant c_7 \tau^r. \tag{1.41}$$

DÉMONSTRATION. Les quantités N_i sont fixées, si bien qu'on est dans la condition

$$\tau_k/\tau_{k+1} \geqslant 1 + d_3, \qquad k = 1, \ldots, s,$$
$$\text{où } d_3 = \min_{1 \leqslant k \leqslant s} (N_{k+1}/N_k) - 1.$$

Les raisonnements suivants sont analogues (aux notations près) à ceux du théorème 3.2, § 1.3.

On rappelle que dans la pratique, on résout, au lieu de (1.9), (1.10), le problème (1.23), (1.24) dont le résidu est évalué par (1.25). On admet que le résidu vient surtout de la résolution imprécise du problème (1.9), (1.10) par la méthode de Newton et on se propose d'établir le critère d'arrêt de ses itérations. Soit δ l'erreur de discrétisation attendue sur la solution corrigée (dans les conditions des théorèmes 1.3, 1.4). Il y a intérêt à procéder par itérations de Newton (1.13), (1.20) de façon que l'erreur de calcul commise soit au plus égale à δ. Pour qu'il en soit ainsi, il suffit d'arrêter le procédé après un pas tel qu'on ait (1.22), où

$$\varepsilon \leqslant c_1 \delta / [(e^{2c_1} - 1) \sum_{k=1}^{s+1} |\gamma_k|].$$

Selon le théorème 1.2, la part de l'erreur de la méthode de Newton dans l'erreur faite en calculant chaque u^{τ_k} est au plus

$$\delta/\sum_{k=1}^{s+1}|\gamma_k|.$$

Il en résulte en raison du § 1.6 que l'erreur de calcul sur la solution corrigée est de l'ordre de δ.

Si l'on a insisté sur l'importance de l'erreur apportée par la méthode de Newton, la cause en est l'impossibilité d'exiger, dans les problèmes susceptibles des procédés itératifs, que l'erreur résultant des itérations soit inférieure aux erreurs d'arrondi ou aux erreurs provenant du calcul des fonctions non arithmétiques. Le caractère accidentel des erreurs de deux derniers types fait qu'un procédé itératif comporte autant de pas qu'on le veut.

2.1.2. Interpolation des développements

Ainsi, la précision voulue est atteinte aux nœuds de ω_τ. S'agissant de trois réseaux ou plus, la méthode présente un inconvénient. Il tient à ce qu'on cherche la solution améliorée aux nœuds communs à tous les réseaux, et que ces points peuvent être peu nombreux. Il se peut en outre qu'on demande la solution approchée en des points qui ne sont en général pas des nœuds. On recourt dans ces cas à l'interpolation par des fonctions splines, par des polynômes trigonométriques, etc. Nous nous placerons dans un cas simple où l'on utilise les polynômes d'interpolation de Lagrange.

On prolonge les fonctions discrètes u^{τ_k} du réseau ω_{τ_k} au segment $[0, 1]$ tout entier. On procède comme suit. On prend un segment élémentaire quelconque $[t_j, t_{j+1}]$ du réseau ω_{τ_k}. La solution u^{τ_k} est définie seulement aux nœuds t_j, t_{j+1}, extrémités du segment. On choisit $r - 2$ nœuds le plus proches de ω_{τ_k} et on définit sur $[t_j, t_{j+1}]$ la quantité $u^{\tau_k}(t)$ égale à la valeur du polynôme d'interpolation de Lagrange basé sur r points ainsi choisis. L'interpolation sur tous les segments élémentaires aboutit à une fonction continue qui est égale à u^{τ_k} aux nœuds de ω_{τ_k}. On la désigne par u^{τ_k}.

Avec les interpolants construits, on calcule une solution approchée de la forme

$$U^H(t) = \sum_{k=1}^{s+1} \gamma_k u^{\tau_k}(t), \qquad t \in [0, 1], \qquad (1.42)$$

les poids γ_k étant déterminés moyennant le système (1.39).

THÉORÈME 1.5. *On suppose que le problème* (1.1), (1.2) *vérifie les conditions* (1.3), (1.4). *La solution corrigée* (1.42) *générée à partir des interpolants* u^{τ_k} *avec les poids donnés par le système* (1.39) *admet l'estimation*

$$\| U^H - u \|_{C[0,1]} \leqslant c_8 \tau^r. \tag{1.43}$$

DÉMONSTRATION. En vertu du théorème 1.3, on a aux nœuds du réseau ω_{τ_k} le développement

$$u^{\tau_k}(t) = u(t) + \sum_{j=1}^{s} \tau_k^{2j} v_j(t) + \tau_k^r \eta^{\tau_k}(t), \tag{1.44}$$

où $v_j \in C^{r-2j+1}[0, 1]$ et ne dépendent pas de τ_k et le reste η^{τ_k} satisfait à

$$\| \eta^{\tau_k} \|_{C,\tau_k} \leqslant c_9. \tag{1.45}$$

On démontre que le développement reste en vigueur pour les prolongements construits des u^{τ_k} à $[0, 1]$ tout entier. Soit $t \in [0, 1]$ quelconque. Il appartient à un segment élémentaire $[t_j, t_{j+1}]$. Nous avons interpolé sur $[t_j, t_{j+1}]$ en $t_i, t_{i+1}, \ldots, t_{i+r-1}$, $i \leqslant j \leqslant i+r-1$; aussi

$$u^{\tau_k}(t) = \sum_{l=0}^{r-1} \alpha_l(t) u^{\tau_k}(t_{i+l}),$$

α_l étant des polynômes de degré $r-1$. Cette formule se récrit en raison de (1.44)

$$u^{\tau_k}(t) = \sum_{l=0}^{r-1} \alpha_l(t) u(t_{i+l}) + \sum_{j=1}^{s} \tau_k^{2j} \sum_{l=0}^{r-1} \alpha_l(t) v_j(t_{i+l}) + \\ + \tau_k^r \sum_{l=0}^{r-1} \alpha_l(t) \eta^{\tau_k}(t_{i+l}). \tag{1.46}$$

On sait que $u \in C^{r+1}[0, 1]$, si bien que la précision de la formule d'interpolation correspondante est en τ_k^r. Aussi

$$\sum_{l=0}^{r-1} \alpha_l(t) u(t_{i+l}) = u(t) + \tau_k^r \rho_0(t), \tag{1.47}$$

où

$$| \rho_0(t) | \leqslant c_{10}.$$

Les fonctions v_j sont moins régulières, et l'interpolation basée sur r points donne une précision plus mauvaise (lemme 3.1, § 7.3):

$$\sum_{l=0}^{r-1} \alpha_l(t) v_j(t_{i+l}) = v_j(t) + \tau_k^{r-2j} \rho_j(t), \tag{1.48}$$

avec

$$|\rho_j(t)| \leqslant c_{11}.$$

On a pour η^{τ_k} une estimation simple, conséquence de (1.45) et de la propriété de borne des coefficients α_l (lemme 3.2, § 7.3):

$$\left|\sum_{l=0}^{r-1} \alpha_l(t)\, \eta^{\tau_k}(t_{i+l})\right| \leqslant c_{12}.$$

On porte dans (1.46) les développements (1.47), (1.48) et on pose

$$\eta^{\tau_k}(t) = \rho_0(t) + \sum_{j=1}^{s} \rho_j(t) + \sum_{l=0}^{r-1} \alpha_l(t)\, \eta^{\tau_k}(t_{i+l}),$$

il vient des développements (1.44) vrais pour [0, 1] tout entier et $k = 1, \ldots, s+1$. Cela étant, on a

$$\left|\eta^{\tau_k}(t)\right| \leqslant c_{10} + s c_{11} + c_{12}. \tag{1.49}$$

On fait la somme de ces développements avec les poids γ_k:

$$U^H(t) = \sum_{k=1}^{s+1} \gamma_k u(t) + \sum_{j=1}^{s} \left(\sum_{k=1}^{s+1} \gamma_k \tau_k^{2j}\right) v_j(t) + \sum_{k=1}^{s+1} \tau_k^r \gamma_k \eta^{\tau_k}(t).$$

Du moment que γ_k satisfont aux équations (1.39), on obtient

$$U^H(t) = u(t) + \sum_{k=1}^{s+1} \tau_k^r \gamma_k \eta^{\tau_k}(t),$$

d'où

$$|U^H(t) - u(t)| \leqslant \sum_{k=1}^{s+1} |\gamma_k|\, |\eta^{\tau_k}(t)|\, \tau_k^r. \tag{1.50}$$

Les pas τ_k vérifient l'égalité

$$\frac{\tau_k}{\tau_{k+1}} = \frac{N_{k+1}}{N_k}.$$

Lorsque $M \to \infty$, les nombres N_k sont fixes, donc

$$\frac{\tau_k}{\tau_{k+1}} \geqslant 1 + c_{13},$$

où

$$c_{13} = \min_{1 \leqslant k \leqslant s} \left(\frac{N_{k+1}}{N_k}\right) - 1 > 0.$$

On peut appliquer le lemme 2.4, § 7.2, d'où la borne des quantités γ_k:

$$|\gamma_k| \leqslant \left(\frac{1 + 2c_{13} + c_{13}^2}{2c_{13} + c_{13}^2}\right)^{s+1}.$$

Comme η^{τ_k} vérifie l'estimation (1.49), il résulte de (1.50):

$$|U^H(t) - u(t)| \leqslant \sum_{k=1}^{s+1} \tau_k^r \left(\frac{1 + 2c_{13} + c_{13}^2}{2c_{13} + c_{13}^2}\right)^{s+1} (c_{10} + c_{11}s + c_{12}).$$

Puisque $\tau_k \geqslant \tau_1 = \tau/N_1$, on a

$$|U^H(t) - u(t)| \leqslant c_8\tau^r,$$

où

$$c_8 = \frac{(s+1)}{N_1^r}\left(\frac{(1 + 2c_{13} + c_{13}^2}{2c_{13} + c_{13}^2}\right)^{s+1} (c_{10} + sc_{11} + c_{12}).$$

Cette inégalité entraîne (1.43), et la constante c_8 est indépendante de t et τ. Le théorème 1.5 se trouve démontré.

2.1.3. Sur le rapport de pas des réseaux

On se demande maintenant quel est, pour les réseaux ω_{τ_k}, le rapport de pas déterminés par $N_1, \ldots, N_{s+1}$ entiers qui garantit la précision maximum. On exige que les données du problème (1.1), (1.2) soient un peu plus régulières:

$$f \in C^{2s+3}([0, 1] \times (-\infty, \infty)).$$

On montre qu'avec cette condition, on connaît plus sur le comportement de la partie principale de l'erreur $U^H - u$. En effet, on a pour les solutions approchées u^{τ_k} par le théorème 1.3

$$u^{\tau_k}(t) = u(t) + \sum_{j=1}^{s+1} \tau_k^{2j} v_j(t) + \xi^{\tau_k}(t), \quad t \in \omega_{\tau_k}, \tag{1.51}$$

où

$$\|\xi^{\tau_k}\|_{C,\tau_k} = O(\tau_k^{2s+3}).$$

On fait la somme avec les poids γ_k vérifiant les équations (1.39), il vient

$$U^H(t) = u(t) + \sum_{k=1}^{s+1} \gamma_k \tau_k^{2s+2} v_{s+1}(t) + \rho^H(t), \; t \in \bar{\omega}_\tau, \qquad (1.52)$$

où

$$\| \rho^H \|_{C,\tau} = O(\tau^{2s+3}).$$

Les quantités v_{s+1} étant indépendantes de τ_k et γ_k, la partie principale de l'erreur est caractérisée en valeur absolue par le coefficient

$$\sum_{k=1}^{s+1} \gamma_k \tau_k^{2s+2}$$

qu'on simplifie moyennant le lemme 2.6, § 7.2. Ce lemme implique

$$\sum_{k=1}^{s+1} \gamma_k \tau_k^{2s+2} = (-1)^{s+2} \prod_{k=1}^{s+1} \tau_k^2.$$

On a déjà dit que le temps de calcul automatique du problème approché (1.9), (1.10) pour un seul réseau $\bar{\omega}_{\tau_k}$ est fonction du nombre d'opérations arithmétiques (portant essentiellement sur le second membre) proportionnel à $1/\tau_k$. Le temps de résolution de $s+1$ problèmes (1.9), (1.10) pour les pas $\tau_1, \ldots, \tau_{s+1}$ est égal presque exactement à

$$\alpha \sum_{k=1}^{s+1} \frac{1}{\tau_k},$$

α étant un temps (en secondes, par exemple) dépendant du type de l'ordinateur, du traducteur du langage algorithmique en langage machine, etc., mais indépendant de τ_k.

On pose le problème suivant: trouver un ensemble d'entiers naturels $N_1, \ldots, N_{s+1}$ tel qu'étant donné le nombre

$$\mu = \alpha \sum_{k=1}^{s+1} \frac{1}{\tau_k} = \frac{\alpha}{\tau} \sum_{k=1}^{s+1} N_k$$

(caractéristique du temps de calcul sur ordinateur de la solution approchée (1.40)), on minimise la quantité

$$\nu = \prod_{k=1}^{s+1} \tau_k^2 = \tau^{2s+2} \prod_{k=1}^{s+1} \frac{1}{N_k^2}$$

(qui caractérise la partie principale de l'erreur sur cette solution).

On utilise la méthode des multiplicateurs de Lagrange et on suppose N_k être des variables continues. On forme la fonction

$$\Phi(N_1, \ldots, N_{s+1}) = \tau^{2s+2} \prod_{k=1}^{s+1} \frac{1}{N_k^2} + \lambda\left(\frac{\alpha}{\tau} \sum_{k=1}^{s+1} N_k - \mu\right)$$

et on cherche la condition nécessaire d'extrémum

$$\frac{\partial \Phi}{\partial N_l} = \frac{-2\tau^{2s+2}}{N_l} \prod_{k=1}^{s+1} \frac{1}{N_k^2} + \lambda \frac{\alpha}{\tau} = 0.$$

Comme α et τ sont positifs, il en est de même de λ, d'où

$$N_1 = \ldots = N_{s+1} = \frac{2\tau^{2s+3}}{\lambda\alpha} \prod_{k=1}^{s+1} \frac{1}{N_k^2}.$$

Ainsi, plus N_i sont voisins, et moins ν est grand pour μ fixé. S'agissant des processus réels, le cas limite $N_1 = \ldots = N_{s+1}$ est à exclure car le système (1.39) devient incompatible et la formule (1.40) n'a plus de sens. Si N_k sont trop rapprochés, γ_k augmente de façon sensible, si bien que l'erreur sur la solution (1.40) risque d'avoir pour partie principale les erreurs d'arrondis dont notre modèle ne tient pas compte.

On étaie ces résultats par les exemples numériques relatifs à plusieurs relations entre N_k.

Le problème

$$u' + tu = (t^2 + t + 1)\, e^t, \qquad t \in (0, 2),$$
$$u(0) = 0 \tag{1.53}$$

possède pour solution la fonction

$$u(t) = te^t.$$

On vérifie aisément qu'on est dans les hypothèses qui permettent d'extrapoler sur deux pas τ_k et plus. On résout d'abord plusieurs problèmes aux différences (1.9), (1.10) avec les pas τ_k, on trouve les erreurs correspondant au point $t = 2$ et on établit la dépendance de l'erreur

$$\zeta(M_k) = \left| u^{\tau_k}(2) - u(2) \right|$$

par rapport au nombre M_k de points du réseau ω_{τ_k}. En coordonnées logarithmiques, ce graphe est donné sur la fig. 2.1.

On forme les solutions améliorées extrapolées sur deux solutions approchées pour divers rapports de pas. La fig. 2.1 visualise trois

relations entre l'erreur sur la solution corrigée au point $t = 2$ et le nombre de nœuds de deux réseaux utilisés. Ces relations correspondent aux trois rapports $N_2 : N_1$ de pas, à savoir 16 : 15, 2 : 1 et 10 : 1. On a travaillé en virgule flottante avec six décimales réservées à la mantisse afin de repérer l'influence des erreurs d'arrondi.

La comparaison des courbes permet d'affirmer que la méthode d'extrapolation est en effet le plus économique dans le cas 16 : 15

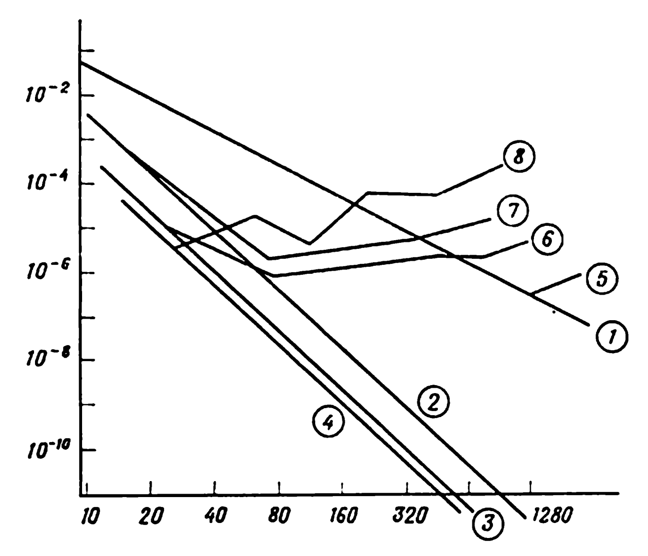


Fig. 2.1. Erreurs sur les solutions approchées du problème (1.53) au point $t = 2$ (avec et sans erreurs d'arrondi)

1 — erreur sur la solution du schéma de Crank-Nicholson (1.9), (1.10) (sans erreurs d'arrondi); *2* — erreur sur la solution extrapolée (1.40) pour le rapport de pas $N_2 : N_1 = 10 : 1$ (sans erreurs d'arrondi); *3* — *ibid.* pour $N_2 : N_1 = 2 : 1$; *4* — *ibid.* pour $N_2 : N_1 = 16 : 15$; *5* — erreur sur la solution du schéma de Crank-Nicholson (1.9), (1.10) (avec erreurs d'arrondi); *6* — erreur sur la solution extrapolée (1.40) pour $N_2 + N_1 = 10 : 1$ (avec erreurs d'arrondi); *7* — *ibid.* pour $N_2 : N_1 = 2 : 1$; *8* — *ibid.* pour $N_2 : N_1 = 16 : 15$.

tant que les erreurs d'arrondi ne surviennent pas. D'autre part, le gain peu important par rapport au cas 2 : 1 ne compense pas ce qu'on perd avec les erreurs d'arrondi dont la répercussion augmente avec le morcellement des pas. On note en outre que les réseaux du second cas possèdent beaucoup plus de points communs, si bien que le coût de l'interpolation (négligé en l'occurrence) du cas 16 : 15 dépasse d'ordinaire l'économie mentionnée. Quant au rapport $N_2 : N_1 = 10 : 1$, l'extrapolation correspondante ne peut être compétitive sur le plan économic avec les deux autres cas.

2.1.4. Extrapolation pour des problèmes rigides

Les résultats de ce paragraphe se transposent aux systèmes d'équations différentielles ordinaires

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}\,(t,\,\mathbf{u}), \qquad t \in (0,\,1), \tag{1.54}$$

$$\mathbf{u}(0) = \mathbf{u}_0,$$

avec $\mathbf{u}_0$, $\mathbf{u}(t)$ des vecteurs de n composantes, $\mathbf{f}$ une fonction vectorielle de $n+1$ variables: $\mathbf{f}(t,\,\mathbf{u}) = \mathbf{f}(t,\,u_1, \ldots,\,u_n)$. On exige que chaque composante f_i de $\mathbf{f}$ satisfasse aux conditions (1.3), (1.4), auquel cas on énonce sans peine le théorème 1.3 en termes de vecteurs.

Voyons comment la méthode d'extrapolation fonctionne pour des systèmes rigides. L'équation différentielle possède dans ce cas des solutions particulières à décroissance exponentielle rapide, et la solution exacte varie modérément sur la plus grande partie de l'intervalle d'intégration. On a surtout affaire à une classe de systèmes (1.54) dont la matrice de Jacobi $J(t)$ d'éléments $J_{ij} = \partial f_i/\partial u_j$, $i, j = 1, \ldots, n$ possède des valeurs propres $\lambda_i(t)$ distinctes telles que

$$\operatorname{Re}(\lambda_i) < 0 \text{ et } \max_{i=1,\ldots,n} \operatorname{Re}(-\lambda_i) \ll 1 \qquad \forall t \in (0,\,1). \tag{1.55}$$

S'agissant des systèmes rigides, les conditions (1.3), (1.4) sont en général vérifiées avec c_1 très grande, si bien que la restriction $\tau \leqslant 1/2\,c_1$ s'avère trop pénible. On justifie toutefois pour les systèmes rigides vérifiant (1.55) la stabilité de la méthode des différences finies (1.9), (1.10) avec $\tau > 0$ quelconque. On démontre ce fait et la convergence de la méthode de Newton par des procédés quelque peu différents pour lesquels nous renvoyons le lecteur à [117] et à la bibliographie de cet ouvrage.

On préfère traiter les problèmes rigides non par la méthode des trapèzes

$$u_{\bar{t}}^{\tau}\,(t) = f_{\tilde{t}}\,(t,\,u^{\tau}\,(t)) \quad \text{sur} \quad \omega_{\tau}, \tag{1.56}$$

$$u^{\tau}\,(0) = u_0,$$

où

$$f_{\tilde{t}}\,(t,\,u\,(t)) = [f\,(t+\tau/2,\,u\,(t+\tau/2)) + f\,(t-\tau/2,\,u\,(t-\tau/2))]/2,$$

mais par la méthode des rectangles implicite. En voici les raisons (voir [108]). 1) On légitime bien pour (1.56) les résultats du n° 2.1.1, mais la recherche de la valeur approchée $u^{\tau}(t+\tau/2)$ par la méthode de Newton est suivie du calcul de $f(t+\tau/2,\,u^{\tau}\,(t+\tau/2))$

qu'on a à utiliser au pas suivant. Aussi la méthode des trapèzes exige plus de calculs du second membre. 2) Soit l'équation modèle

$$u' = \lambda(t)\, u, \tag{1.57}$$

où $\lambda(t) < 0$, si bien que $u(t)$ diminue avec la croissance de t. La méthode des trapèzes entraîne

$$u^\tau(t + \tau/2) = \frac{1 + \tau\lambda(t - \tau/2)/2}{1 - \tau\lambda(t + \tau/2)/2}\, u^\tau(t - \tau/2).$$

L'inégalité $|u^\tau(t + \tau/2)| \leqslant |u^\tau(t - \tau/2)|$ a donc lieu sous la condition nécessaire

$$\tau(\lambda(t + \tau/2) - \lambda(t - \tau/2)) \leqslant 4.$$

Si la fonction $\lambda(t)$ croît sur une portion de l'intervalle d'intégration, cette condition conduit à une contrainte relative à la longueur du pas. Cette contrainte n'intervient pas dans la méthode implicite des rectangles encore que le problème d'approcher les solutions à amortissement rapide se pose toujours.

On tourne la dernière difficulté en posant λ de (1.57) égale à une constante complexe telle que $\mathrm{Re}(\lambda) < 0$. La méthode des rectangles entraîne

$$u^\tau(t + \tau/2) = \frac{1 + \tau\lambda/2}{1 - \tau\lambda/2}\, u^\tau(t - \tau/2). \tag{1.58}$$

Si

$$\mathrm{Re}(\lambda\tau) \ll 0, \tag{1.59}$$

la solution approchée $u^\tau(t)$ est approximativement égale à $(-1)^{t/\tau}u(0)$. On obtient une solution approchée oscillante et non la solution $|u(t)| = e^{\mathrm{Re}(\lambda)t}$ qui s'amortit rapidement. On évite cet inconvénient par l'artifice proposé par Lindberg [107] qui consiste à lisser la solution approchée par la formule

$$\tilde{u}^\tau(t) = u^\tau_{\bar{t}t}(t), \tag{1.60}$$

où $u_{\bar{t}t}(t) = [u(t - \tau) + 2\, u(t) + u(t + \tau)]/4$. L'égalité (1.58) implique

$$\tilde{u}^\tau(t) = \frac{1}{1 - (\tau\lambda/2)^2}\, u^\tau(t),$$

et la solution est atténuée par suite de (1.59). Si l'on applique (1.60) plusieurs (M) fois et si l'on initialise les approximations avec le résultat de lissage, on a

$$\tilde{u}^\tau(t) = \frac{1}{(1 - (\tau\lambda/2)^2)^M}\, u^\tau(t).$$

Ainsi, le lissage répété un nombre suffisant de fois permet d'étouffer les termes correspondants de la solution approchée. Dans le cas du système rigide (1.54), la solution renferme aussi bien des termes à amortissement rapide que des composantes à variation lente. S'agissant de ces dernières, le lissage (1.60) conserve le développement (1.29). ainsi que les résultats correspondants relatifs à la précision de l'extrapolation. L'ouvrage [117] situe la méthode des rectangles implicite parmi les procédés de calcul et en décrit plusieurs variantes.

2.2. Schémas aux différences explicites

Nous nous proposons d'examiner dans ce paragraphe le problème de Cauchy pour une équation différentielle ordinaire non linéaire du premier ordre. La résolution approchée des équations non linéaires est très actuelle. Elle est effectuée d'ordinaire par l'algorithme de Runge-Kutta ou par des méthodes linéaires à pas liés. Il est également possible d'atteindre une précision élevée si l'on extrapole les solutions discrètes simples sur le pas du réseau. C'est ce dernier procédé que nous voulons étudier.

2.2.1. Méthode d'Euler

La technique d'Euler occupe une place à part parmi les méthodes numériques de résolution des problèmes non linéaires. Elle opère par approximations explicites d'équations, et sa mise en œuvre particulièrement simple fait oublier son défaut d'être exacte à l'ordre un.

Soit l'équation

$$\frac{du}{dt} = f(t, u), \qquad t \in (0, 1), \tag{2.1}$$

avec la condition initiale

$$u(0) = u_0, \tag{2.2}$$

$f(t, u)$ étant une fonction réelle et

$$f \in C^r([0, 1] \times (-\infty, \infty)), \; r \geqslant 2. \tag{2.3}$$

On suppose que le problème admet une solution unique u et que

$$u \in C^{r+1}[0, 1]. \tag{2.4}$$

On utilise la méthode d'Euler sur un réseau uniforme de pas $\tau = 1/M$:

$$u_{\bar{t}}^{\tau} = f(t, u^{\tau}), \qquad t \in \hat{\omega}_{\tau}, \tag{2.5}$$

$$u^{\tau}(0) = u_0. \tag{2.6}$$

Le problème aux différences ainsi obtenu est non linéaire, si bien qu'on ne saurait appliquer directement les théorèmes généraux du Chapitre premier. Quant à justifier la précision élevée de la solution extrapolée, on procède comme plus haut.

Connaissant la solution $u(t)$ du problème (2.1), (2.2), on construit le système d'équations

$$\sum_{s=1}^{l+1} \frac{1}{s!} \frac{d^s v_{l-s+1}}{dt^s} = \sum_{s=1}^{l} \frac{1}{s!} \frac{\partial^s f(t, u)}{\partial u^s} \sum_{i_1+\dots+i_s=l} v_{i_1} \dots v_{i_s},$$

$$t \in (0, 1)\ *, \quad (2.7)$$

avec les conditions initiales pour les fonctions inconnues $v_1, \dots, v_{r-1}$:

$$v_l(0) = 0, \qquad l = 1, \dots, r-1. \qquad (2.8)$$

On pose $v_0 = u$ et on cherche v_l successifs dans l'ordre de croissance de l. En effet, l'équation (2.7) s'écrit pour $l = 1$:

$$\frac{dv_1}{dt} - v_1 \frac{\partial f(t, u)}{\partial u} = -\frac{1}{2} \frac{d^2 v_0}{dt^2}, \qquad t \in (0, 1). \qquad (2.9)$$

Il est évident qu'il s'agit, pour $v_0 = u$ connue, d'une équation linéaire par rapport à v_1. Comme les coefficients de (2.1) sont de classe C^{r-1} [0, 1], il existe par suite de la linéarité une solution unique v_1 de C^r [0, 1] avec les conditions initiales $v_1(0) = 0$. Les fonctions v_l suivantes sont définies de façon analogue. On suppose définie la l-1-ième fonction $v_{l-1} \in C^{r+2-l}$ [0, 1]. On récrit (2.7):

$$\frac{dv_l}{dt} - v_l \frac{\partial f}{\partial u}(t, u) = \sum_{s=2}^{l} \frac{1}{s!} \frac{\partial^s f(t, u)}{\partial u^s} \sum_{i_1+\dots+i_s=l} v_{i_1} \dots v_{i_s} -$$

$$- \sum_{s=2}^{l+1} \frac{1}{s!} \frac{d^s v_{l-s+1}}{dt^s}, \qquad t \in (0, 1). \quad (2.10)$$

Il est immédiat de vérifier que le plus grand indice des fonctions v_k du second membre est au plus $l-1$ et que chaque terme de ce membre est au moins $r - l$ fois continûment dérivable sur [0, 1]. Aussi la solution de l'équation linéaire (2.10) avec la condition initiale homogène $v_l(0) = 0$ existe, est unique et appartient à C^{r-l+1} [0, 1].

---

* Ici $i_1, \dots, i_s$ sont des indices entiers positifs et $\sum\limits_{i_1+\dots+i_s=l}$ signifie la sommation par rapport à toutes les combinaisons d'indices dont la somme vaut l; si $s > l$, alors la somme est estimée être nulle.

Connaissant v_l et la solution du problème (2.5), (2.6), on construit la fonction discrète

$$\eta^\tau = \tau^{-r}\left(u^\tau - \sum_{l=0}^{r-1} \tau^l v_l\right) \quad \text{sur} \quad \bar\omega_\tau. \tag{2.11}$$

THÉORÈME 2.1. *Si l'on est dans les conditions* (2.3), (2.4), *la solution du problème aux différences* (2.5), (2.6) *admet le développement*

$$u^\tau = u + \sum_{l=1}^{r-1} \tau^l v_l + \tau^r \eta^\tau \quad \textit{sur} \quad \bar\omega_\tau, \tag{2.12}$$

avec v_l *définies à partir du système* (2.7), (2.8) *et indépendantes de* τ. *Cela étant, si les solutions approchées sont bornées uniformément en* τ *:

$$\| u^\tau \|_{C,\tau} \leqslant c_2, \tag{2.13}$$

le reste η^τ *est également borné :*

$$\| \eta^\tau \|_{C,\tau} \leqslant c_3. \tag{2.14}$$

DÉMONSTRATION. On remplace u^τ de (2.5) par ses valeurs (2.12). Vu que $v_0 = u$,

$$\sum_{l=0}^{r-1} \tau^l (v_l)_t + \tau^r \eta_t^\tau = f\left(t, \sum_{l=0}^{r-1} \tau^l v_l + \tau^r \eta^\tau\right).$$

On transforme le premier membre par le lemme 1.1, § 7.1, et on développe le second membre en formule de Taylor par rapport à la deuxième variable de la fonction $f(t, u)$:

$$\sum_{l=0}^{r-1} \tau^l \left(\sum_{s=0}^{r-l-1} \tau^s \frac{1}{(s+1)!} \frac{d^{s+1} v_l}{dt^{s+1}} + \tau^{r-l} \rho_{r-l}^\tau\right) + \tau^r \eta_t^\tau =$$
$$= f\left(t, \sum_{l=0}^{r-1} \tau^l v_l\right) + \tau^r \eta^r \sigma^\tau. \tag{2.15}$$

Ici

$$\sigma^\tau = \frac{\partial f}{\partial u}(t, \xi^\tau).$$

---

* Voici une condition qui donne (2.13) en raison de [65]:

$$\max_{t \in [0,1]} \sup_{u \in (-\infty, \infty)} \left| \frac{df}{du}(t, u) \right| \leqslant c_1.$$

avec ξ^τ un point de l'intervalle d'extrémités

$$\sum_{l=0}^{r-1} \tau^l v_l, \ u^\tau.$$

On note que la continuité de

$$\sum_{l=0}^{r-1} \tau^l v_l$$

sur $[0, 1]$ et la borne uniforme de la fonction discrète u^τ sur $\bar{\omega}_\tau$ font que toutes les valeurs de ξ^τ appartiennent à un segment fini $[-c_4, c_4]$ quels que soient $\tau > 0$ et $t \in \hat{\omega}_\tau$. La continuité de la dérivée $\frac{\partial f}{\partial u}(t, u)$ sur le rectangle $[0, 1] \times [-c_4, c_4]$ en entraîne la borne, si bien que

$$|\sigma^\tau(t)| \leqslant c_5 \qquad \forall t \in \hat{\omega}_\tau. \tag{2.16}$$

Les coefficients ρ^τ_{r-l} des restes sont bornés:

$$\sum_{l=0}^{r-1} |\rho^\tau_{r-l}(t)| \leqslant c_6 \qquad \forall t \in \hat{\omega}_\tau. \tag{2.17}$$

vu la régularité des fonctions v_l.

On intervertit l'ordre de sommation dans le premier membre de (2.15), et on applique au second membre la formule de Taylor, il vient

$$\sum_{l=0}^{r-1} \tau^l \sum_{s=1}^{l+1} \frac{1}{s!} \frac{d^s v_{l-s+1}}{dt^s} + \tau^r \sum_{l=0}^{r-1} \rho^\tau_{r-l} + \tau^r \eta^\tau_t = f(t, u) +$$
$$+ \sum_{l=1}^{r-1} \left\{ \sum_{s=1}^{r-1} \tau^s v_s \right\}^l \frac{1}{l!} \frac{\partial^l f}{\partial u^l}(t, u) + \tau^r g^\tau + \tau^r \eta^\tau \sigma^\tau. \tag{2.18}$$

Ici

$$g^\tau = \frac{1}{r!} \left\{ \sum_{s=1}^{r-1} \tau^{s-1} v_s \right\}^r \frac{\partial^r f}{\partial u^r}(t, \zeta^\tau),$$

avec ζ^τ un point du segment d'extrémités

$$u, \quad \sum_{l=0}^{r-1} \tau^l v_l.$$

Les fonctions v_i étant continues sur [0, 1] sont bornées. Aussi la quantité ζ^τ est uniformément bornée:

$$|\zeta^\tau(t)| \leqslant c_7 \qquad \forall t \in \hat{\omega}_\tau. \tag{2.19}$$

La propriété de

$$\frac{\partial^r f}{\partial u^r}(t, u)$$

d'être continue sur le rectangle $[0, 1] \times [-c_7, c_7]$ implique la borne de g^τ:

$$|g^\tau(t)| \leqslant c_8 \qquad \forall t \in \hat{\omega}_\tau. \tag{2.20}$$

On transforme la somme double dans le second membre de (2.18). On élève à la puissance les accolades et on identifie les coefficients des mêmes puissances de τ:

$$\sum_{l=1}^{(r-1)^2} \tau^l \sum_{s=1}^{r-1} \frac{1}{s!} \frac{\partial^s f(t, u)}{\partial u^s} \sum_{i_1+\ldots+i_s=l} v_{i_1} \ldots v_{i_s}.$$

On réunit tous les termes sauf les termes en τ^l, $l < r$:

$$\tau^r b^\tau = \sum_{l=r}^{(r-1)^2} \tau^l \sum_{s=1}^{r-1} \frac{1}{s!} \frac{\partial^s f(t, u)}{\partial u^s} \sum_{i_1+\ldots+i_s=l} v_{i_1} \ldots v_{i_s}.$$

On divise membre à membre par τ^r et on passe aux modules sans oublier que $\tau \leqslant 1$, il vient l'estimation

$$|b^\tau| \leqslant \sum_{l=r}^{(r-1)^2} \sum_{s=1}^{r-1} \frac{1}{s!} \left| \frac{\partial^s f(t, u)}{\partial u^s} \right| \sum_{i_1+\ldots+i_s=l} |v_{i_1} \ldots v_{i_s}|.$$

Comme les fonctions $u(t)$, $v_i(t)$, $\dfrac{\partial^s f(t, u(t))}{\partial u^s}$ sont indépendantes de τ et uniformément bornées sur [0, 1], l'inégalité se récrit

$$|b^\tau| \leqslant c_9 \qquad \forall t \in \hat{\omega}_\tau. \tag{2.21}$$

Ainsi, on a mis le second membre de l'égalité (2.18) sous la forme

$$f(t, u) + \sum_{l=1}^{r-1} \tau^l \sum_{s=1}^{l} \frac{1}{s!} \frac{\partial^s f(t, u)}{\partial u^s} \sum_{i_1+\ldots+i_s=l} v_{i_1} \ldots v_{i_s} + \tau^r b^\tau +$$
$$+ \tau^r g^\tau + \tau^r \eta^\tau \sigma^\tau.$$

On utilise le système (2.7) et les expressions résultant des transformations effectuées sur les deux membres de (2.18). On a

$$\tau^r \sum_{l=0}^{r-1} \rho_{r-l}^\tau + \tau^r \eta_{\bar{t}}^\tau = \tau^r b^\tau + \tau^r g^\tau + \tau^r \eta^\tau \sigma^\tau \quad \text{sur} \quad \dot{\omega}_\tau,$$

ou

$$\eta_{\bar{t}}^\tau - \sigma^\tau \eta^\tau = b^\tau + g^\tau - \sum_{l=0}^{r-1} \rho_{r-l}^\tau \quad \text{sur} \quad \dot{\omega}_\tau. \tag{2.22}$$

Les estimations (2.16), (2.17), (2.20), (2.21) aidant, on en tire

$$|\eta^\tau(t+\tau)| \leqslant |\eta^\tau(t)|\,(1+\tau c_5) + \tau(c_6+c_8+c_9), \quad t \in \dot{\omega}_\tau. \tag{2.23}$$

Nous allons nous servir du résultat ci-dessous qu'on établit sans peine par la méthode de récurrence (voir par exemple [52]).

Lemme 2.2. *On suppose que la fonction discrète* ξ *est définie sur* $\bar{\omega}_\tau$ *et vérifie l'inégalité*

$$|\xi(t+\tau)| \leqslant |\xi(t)|\,(1+\tau\delta) + \tau B, \qquad t \in \dot{\omega}_\tau,$$

avec $B \geqslant 0$ *et* $\delta > 0$. *Alors*

$$|\xi(t)| \leqslant e^{\delta t}|\xi(0)| + \frac{e^{\delta t}-1}{\delta} B, \qquad t \in \bar{\omega}_\tau.$$

La valeur initiale $\eta^\tau(0)$ est obtenue à partir de la définition de la fonction η^τ et des égalités (2.2), (2.6), (2.8):

$$\eta^\tau(0) = \tau^{-r}\left(u^\tau(0) - \sum_{l=0}^{r-1} \tau^l v_l(0) \right) = 0.$$

Aussi l'inégalité (2.23) et le lemme 2.2 entraînent l'estimation

$$|\eta^\tau(t)| \leqslant \frac{e^{c_5 t}-1}{c_5}(c_6+c_8+c_9), \qquad t \in \bar{\omega}_\tau.$$

On pose $c_3 = (e^{c_5}-1)(c_6+c_8+c_9)/c_5$. On a en raison de $t \leqslant 1$

$$\|\eta^\tau\|_{C,\tau} \leqslant c_3, \tag{2.24}$$

ce qui démontre le théorème 2.1.

On améliore les solutions approchées à l'aide du développement (2.12).

On suppose vérifiées les conditions (2.3), (2.4) et on construit pour $0 < N_1 < \ldots < N_r$ entiers fixés les réseaux ω_{τ_k} de pas $\tau_k = 1/(N_k M)$, avec M un entier naturel qui croît indéfiniment. On résout sur chaque ω_{τ_k} le problème aux différences (2.5), (2.6). Toutes les solutions u^{τ_k} sont définies sur le réseau ω_τ de pas $\tau = 1/M$.

Soit le système

$$\begin{gathered} \sum_{k=1}^{r} \gamma_k = 1, \\ \sum_{k=1}^{r} \gamma_k \tau_k^j = 0, \qquad j = 1, \ldots, r-1. \end{gathered} \tag{2.25}$$

τ_k sont distincts deux à deux, si bien que le déterminant du système n'est pas nul et il existe une solution unique $\gamma_1, \ldots, \gamma_k$. On forme une combinaison linéaire avec ces poids, à savoir

$$U^H(t) = \sum_{k=1}^{r} \gamma_k u^{\tau_k}(t), \qquad t \in \omega_\tau. \tag{2.26}$$

La solution U^H est exacte à l'ordre r en τ tandis que la précision de chaque u^{τ_k} est $O(\tau^1)$.

Théorème 2.3. *On suppose qu'on est, pour le problème* (2.1), (2.2), *dans les conditions* (2.3), (2.4) *avec* $r \geqslant 1$ *entier. La solution corrigée* (2.26) *avec les poids obtenus à partir du système* (2.25) *admet la majoration*

$$\| U^H - u \|_{C, \tau} \leqslant c_{10} \tau^r. \tag{2.27}$$

Démonstration. D'après le théorème 2.1, on a en chaque nœud du réseau ω_τ les développements

$$u^{\tau_k} = u + \sum_{l=1}^{r-1} \tau_k^l v_l + \tau_k^r \eta^{\tau_k} \quad \text{sur} \quad \omega_\tau. \tag{2.28}$$

On les additionne avec les poids γ_k:

$$U^H = u \sum_{k=1}^{r} \gamma_k \tau_k^r \eta^{\tau_k} \quad \text{sur} \quad \omega_\tau. \tag{2.29}$$

L'estimation (2.14) entraîne celle de $|\eta^{\tau k}|$, et le lemme 2.3, § 7.2 celle de $|\gamma_k|$. Aussi l'affirmation du théorème découle de (2.29).

Si l'on interpole moyennant les polynômes de Lagrange, on élabore un algorithme de raffinement pour les points qui ne sont pas communs à tous les réseaux à la fois. Ce procédé est décrit en détail plus haut (voir § 2.1).

2.2.2. Un schéma aux différences centrales explicite

On note qu'avec le schéma d'Euler, le développement de la solution approchée suivant les puissances de τ renferme des puissances impaires. On élimine chaque terme de sa partie régulière en résolvant un problème aux différences auxiliaire de la forme (2.5), (2.6). Il y a donc intérêt à construire une méthode explicite telle que la partie régulière dudit développement ne contienne que les puissances paires de τ.

Soit le schéma aux différences centrales (qu'on appelle également la méthode des rectangles explicite)

$$u^{\tau}_{\mathring{t}} = f(t, u^{\tau}), \qquad t \in \omega_{\tau}, \tag{2.30}$$

où

$$u^{\tau}_{\mathring{t}}(t) = [u^{\tau}(t+\tau) - u^{\tau}(t-\tau)]/2\tau.$$

L'équation (2.30) renferme pour chaque t les valeurs prises par u^{τ} en trois nœuds, si bien que la recherche des valeurs successives u^{τ} sous forme explicite exige deux conditions initiales. La première condition découle de (2.2):

$$u^{\tau}(0) = u_0, \tag{2.31}$$

et la seconde doit garantir la propriété « le développement de la solution approchée u^{τ} renferme seulement les puissances paires de τ ». Cette condition est énoncée par Gragg (voir [91]):

$$u^{\tau}(\tau) = u_0 + \tau f(0, u_0). \tag{2.32}$$

On trouve à partir de l'équation (2.30), $t = \tau, 2\tau, \ldots, s$, sous les conditions (2.31), (2.32) les valeurs de la fonction $u^{\tau}(t)$ pour $t = 0, \tau, \ldots, 1+\tau$. On note qu'on a pu prendre pour deux valeurs initiales les valeurs aux points $-\tau$ et 0. Il y a plus. Si l'on pose

$$u^{\tau}(-\tau) = u_0 - \tau f(0, u_0), \qquad u^{\tau}(0) = u_0, \tag{2.33}$$

alors les équations (2.30) donnent la même solution aux différences parce que (2.33) et (2.30) entraînent pour $t = 0$ l'égalité (2.32):

$$u^\tau(\tau) = u^\tau(-\tau) + 2\tau f(0, u^\tau(0)) = u_0 + \tau f(0, u_0).$$

On suppose qu'on est pour le problème (2.1), (2.2) dans les conditions (2.3), (2.4), avec $r \geqslant 2$ entier. On pose $s = [(r-1)/2]$. Connaissant la solution u, on écrit le système linéaire pour $l = 1, 2, \dots, s$:

$$\begin{aligned} v_l' - \frac{\partial f}{\partial u}(t, u)\, w_l &= -\frac{u^{(2l+1)}}{(2l+1)!} - \sum_{k=1}^{l-1} \frac{v_{l-k}^{(2k+1)}}{(2k+1)!} + \\ &+ \sum_{k=2}^{l} \frac{1}{k!} \frac{\partial^k f}{\partial u^k}(t, u) \sum_{i_1+\dots+i_k=l} w_{i_1} \dots w_{i_k}, \\ w_l' - \frac{\partial f}{\partial u}(t, u)\, v_l &= -\frac{u^{(2l+1)}}{(2l+1)!} - \sum_{k=1}^{l-1} \frac{w_{l-k}^{(2k+1)}}{(2k+1)!} + \\ &+ \sum_{k=2}^{l} \frac{1}{k!} \frac{\partial^k f}{\partial u^k}(t, u) \sum_{i_1+\dots+i_k=l} v_{i_1} \dots v_{i_k} \quad \forall t \in (0, 1). \end{aligned} \tag{2.34}$$

On prend pour conditions initiales

$$v_l(0) = -\frac{1}{(2l)!} u^{(2l)}(0) - \sum_{k=1}^{l-1} \frac{1}{(2k)!} v_{l-k}^{(2k)}(0), \tag{2.35}$$

$$w_l(0) = 0. \tag{2.36}$$

Ce système fournit de proche en proche v_l, w_l dans l'ordre de croissance de l. On a par exemple pour $l = 1$ le système

$$\begin{aligned} v_1' - \frac{\partial f}{\partial u}(t, u)\, w_1 &= -\frac{u'''}{6}, \\ w_1' - \frac{\partial f}{\partial u}(t, u)\, v_1 &= -u'''/6 \qquad \forall t \in (0, 1), \end{aligned} \tag{2.37}$$

avec les conditions initiales

$$v_1(0) = -u''(0)/2, \qquad w_1(0) = 0. \tag{2.38}$$

Les coefficients et le second membre du système (2.37) sont de classe $C^{r-2}[0, 1]$, si bien que la linéarité des équations implique l'existence d'un seul couple de fonctions $v_1, w_1 \in C^{r-1}[0, 1]$ vérifiant les conditions (2.37), (2.38). On procède de même pour les fonctions suivantes. On suppose, par exemple, connues $v_k, w_k \in C^{r+1-2k}[0, 1]$,

$k = 1, \ldots, l-1$. Il est immédiat de vérifier que les indices des fonctions v_k, w_k des seconds membres de (2.34) et de (2.35) sont au plus égaux à $l-1$. Tous les termes du second membre des équations (2.34) sont de plus au moins $r - 2l$ fois continûment dérivables. Il existe donc une solution unique

$$v_l,\ w_l \in C^{r+1-2l}[0,\ 1].$$

On prolonge u, v_l, w_l par τ hors de $[0, 1]$ en des fonctions de même classe. On utilise à cet effet la série de Taylor de longueur correspondante. Par exemple,

$$u(t) = \sum_{k=0}^{r+1} \frac{t^k}{k!} u^{(k)}(0) \qquad \forall\, t \in [-\tau,\ 0],$$

$$u(t) = \sum_{k=0}^{r+1} \frac{(t-1)^k}{k!} u^{(k)}(1) \qquad \forall\, t \in [1,\ 1+\tau].$$

THÉORÈME 2.4. *Si l'on est dans les conditions* (2.3), (2.4), *alors la solution du problème aux différences* (2.33), (2.30) *admet les développements*

$$u^\tau = u + \sum_{l=1}^{s} \tau^{2l} v_l + \eta^\tau, \qquad t = -\tau, \tau, 3\tau, \ldots, \tag{2.39}$$

$$u^\tau = u + \sum_{l=1}^{s} \tau^{2l} w_l + \eta^\tau, \qquad t = 0,\ 2\tau,\ 4\tau, \ldots \tag{2.40}$$

Ici $s = [(r - 1/2)]$, *les fonctions* v_l, w_l *sont définies par le système* (2.34) *à* (2.36) *et ne dépendent pas de* τ. *Cela étant, si les solutions approchées* u^τ *sont bornées uniformément en* τ:

$$\| u^\tau \|_{C,\tau} \leqslant c_{12}, \tag{2.41}$$

alors le reste η^τ *admet l'estimation*

$$\| \eta^\tau \|_{C,\tau} \leqslant c_{13} \tau^r. \tag{2.42}$$

DÉMONSTRATION. Les fonctions u^τ, u, v_l, w_l sont données à l'avance, si bien que les développements (2.39), (2.40) définissent η^τ de façon unique aux nœuds $-\tau, 0, \ldots, 1+\tau$. Il reste à démontrer la validité de (2.42). On porte (2.39), (2.40) dans l'équation (2.30). Soit d'abord le cas t/τ pair, où $t \in \bar{\omega}_\tau$. On a

$$u_{\bar{t}t} + \sum_{l=1}^{s} \tau^{2l} (v_l)_{\bar{t}t} + \eta^\tau_{\bar{t}t} = f\left(t,\ u + \sum_{l=1}^{s} \tau^{2l} w_l + \eta^\tau\right).$$

On transforme le premier membre par le lemme 1.1, § 7.1, et on développe le second membre en formule de Taylor:

$$\sum_{k=0}^{s} \frac{\tau^{2k}}{(2k+1)!} u^{(2k+1)} + \tau^{r} \rho_0^{\tau} + \sum_{l=1}^{s} \tau^{2l} \left[\sum_{k=0}^{s-l} \frac{\tau^{2k}}{(2k+1)!} \times v_l^{(2k+1)} + \tau^{r-2l} \rho_l^{\tau} \right] + \eta_{\bar{t}}^{\tau} = f(t, u) + \sum_{l=1}^{s} \tau^{2l} w_l + \eta^{\tau} \sigma^{\tau}. \quad (2.43)$$

Ici $\sigma^{\tau} = \frac{\partial f}{\partial u} (t, \zeta^{\tau})$, avec ζ^{τ} dans l'intervalle d'extrémités

$$u^{\tau}, \qquad u + \sum_{l=1}^{s} \tau^{2l} w_l .$$

Etant donnée la régularité des fonctions u, w_l, v_l, f, on a

$$|\sigma^{\tau}| \leqslant c_{14}, \quad \sum_{l=0}^{1} |\rho_l^{\tau}| \leqslant c_{15} \quad \text{pour} \quad t = 0,\ 2\tau, \ldots \quad (2.44)$$

On intervertit l'ordre de sommation dans le premier membre de (2.43), tandis que le second membre est une fois de plus développée en formule de Taylor:

$$u' + \sum_{l=1}^{s} \tau^{2l} \left[\frac{u^{(2l+1)}}{(2l+1)!} + \sum_{k=0}^{l-1} \frac{v_{l-k}^{(2k+1)}}{(2k+1)!} \right] + \tau^{r} \sum_{l=0}^{s} \rho_l^{\tau} + \eta_{\bar{t}}^{\tau} = f(t, u) + \sum_{k=1}^{s} \frac{1}{k!} \left(\sum_{l=1}^{s} \tau^{2l} w_l \right)^{k} \times \frac{\partial^k f}{\partial u^k} (t, u) + \tau^{r} \zeta_1^{\tau} + \eta^{\tau} \sigma^{\tau}. \quad (2.45)$$

On a pour ζ_1^{τ}

$$|\zeta_1^{\tau}| \leqslant c_{16}, \quad (2.46)$$

conséquence de la borne des fonctions w_l et de la continuité de la dérivée $\frac{\partial^{s+1} f}{\partial u^{s+1}}$. On effectue l'opération puissance dans la somme double du second membre de (2.45) et on identifie les termes contenant les mêmes puissances de τ, il vient

$$\sum_{l=1}^{s^2} \tau^{2l} \sum_{k=1}^{s} \frac{1}{k!} \frac{\partial^k f}{\partial u^k} (t, u) \sum_{i_1 + \ldots + i_k = l} w_{i_1} \cdots w_{i_k} .$$

On conserve les termes en $\tau^2, \tau^4, \ldots, \tau^{2s}$, et on récrit les termes restants:

$$\zeta_2^\tau = \sum_{l=s+1}^{s^2} \tau^{2l} \sum_{k=1}^{l} \frac{1}{k!} \frac{\partial^k f}{\partial u^k}(t, u) \sum_{i_1+\ldots+i_k=l} w_{i_1} \ldots w_{i_k}.$$

Il est clair que

$$|\zeta_2^\tau| \leqslant \tau^r c_{17}. \tag{2.47}$$

Ainsi, l'équation (2.45) se ramène à

$$u' + \sum_{l=1}^{1} \tau^{2l} \left(\frac{u^{(2l+1)}}{(2l+1)!} + \sum_{k=0}^{l-1} \frac{v_{l-k}^{(2k+1)}}{(2k+1)!} + \right.$$

$$+ \tau^r \sum_{l=0}^{1} \rho_l^\tau + \eta_{tt}^\tau = f(t, u) + \sum_{l=1}^{s} \tau^{2l} \times$$

$$\times \sum_{k=1}^{l} \frac{1}{k!} \frac{\partial^k f}{\partial u^k}(t, u) \sum_{i_1+\ldots+i_k=l} w_{i_1} \ldots w_{i_k} + \tau^r \zeta_1^\tau + \tau^r \zeta_2^\tau + \sigma^\tau \eta^\tau.$$

On effectue une réduction des termes moyennant l'équation (2.1) et le système (2.34), il vient l'egalité

$$\eta_{tt}^\tau - \sigma^\tau \eta = \tau^r \zeta_3^\tau \tag{2.48}$$

dont le second membre admet, en vertu de (2.44), (2.46), (2.47), l'estimation

$$|\zeta_3^\tau| \leqslant c_{18} = c_{15} + c_{16} + c_{17}. \tag{2.49}$$

On obtient de même l'équation (2.48) et l'estimation (2.49) pour $t = \tau, 3\tau, \ldots$, si bien qu'elles sont vérifiées pour tous les $t \in \omega_\tau$.

Quelles sont les conditions initiales pour le système (2.48)? On a pour $t = 0$

$$\eta^\tau(0) = u^\tau(0) - u(0) - \sum_{l=1}^{s} \tau^{2l} w_l(0).$$

Avec les conditions initiales (2.2), (2.31), (2.36), on en déduit

$$\eta^\tau(0) = 0. \tag{2.50}$$

On a pour $t \pm \tau$ les développements

$$\eta^\tau = u^\tau - u - \sum_{l=1}^{s} \tau^{2l} v_l.$$

On passe à la demi-somme. Etant données les conditions (2.32), (2.33),

$$[\eta^\tau(\tau) + \eta^\tau(-\tau)]/2 = u_0 - [u(\tau) + u(-\tau)]/2 - \\ - \sum_{l=1}^{s} \tau^{2l}[v_l(\tau) + v_l(-\tau)]/2.$$

On transforme le second membre par le lemme 1.1, § 7.1 :

$$[\eta^\tau(\tau) + \eta^\tau(-\tau)]/2 = u_0 - \sum_{l=0}^{s} \tau^{2l} \frac{u^{(2l)}(0)}{(2l)!} + \\ + \sum_{l=1}^{s} \tau^{2l} \sum_{k=1}^{s-l} \frac{\tau^{2k}}{(2k)!} v_l^{(2k)}(0) + \tau^r\zeta_4^\tau = - \sum_{l=1}^{s} \tau^{2l} \left[\frac{u^{(2l)}(0)}{(2l)!} + \right. \\ \left. + \sum_{k=0}^{l-1} \frac{v_{l-k}^{(2k)}(0)}{(2k)!} \right] + \tau^r\zeta_4^\tau.$$

La constante ζ_4^τ satisfait à l'inégalité

$$|\zeta_4^\tau| \leqslant c_{19}. \tag{2.51}$$

On a, compte tenu de (2.35),

$$[\eta^\tau(\tau) + \eta^\tau(-\tau)]/2 = \tau^r\zeta_4^\tau. \tag{2.52}$$

L'égalité (2.48) entraîne pour $t = 0$

$$[\eta^\tau(\tau) - \eta^\tau(-\tau)]/2\tau = \sigma^\tau(0)\,\eta^\tau(0) + \tau^r\zeta_3^\tau(0).$$

On multiplie par $+\tau$ (ou par $-\tau$) et on additionne avec (2.52). On a par suite de (2.50)

$$\eta^\tau(\pm\tau) = \tau^r\xi_5^\tau, \quad \text{où} \quad |\zeta_5^\tau| \leqslant c_{20}. \tag{2.53}$$

On évalue les autres valeurs de η^τ à l'aide du

Lemme 2.5. *On suppose que la fonction discrète ζ vérifie l'inégalité* $|\zeta(t + 2\tau)| \leqslant |\zeta(t)| + |\zeta(t+\tau)|\,\tau\delta + \tau B$ $\forall t = 0, \tau, 2\tau, \ldots$, *où* $B \geqslant 0$, $\delta \geqslant 0$. *On a*

$$|\zeta(t)| \leqslant e^{\delta t}(|\zeta(0)| + |\zeta(\tau)| + tB). \tag{2.54}$$

Démonstration. Le résultat est évident pour $t = 0, \tau$. On le suppose vrai pour certains t, $t+\tau$, et on l'établit pour $t + 2\tau$. On a

$$|\zeta(t + 2\tau)| \leqslant e^{\delta t}(|\zeta(0)| + |\zeta(\tau)| + tB + e^{\delta(t+\tau)}(|\zeta(0)| + \\ + |\zeta(\tau)| + (t+\tau)B)\tau\delta + \tau B \leqslant e^{\delta(t+2\tau)}(|\zeta(0)| + |\zeta(\tau)|) + \\ + (t+\tau)Be^{\delta(t+\tau)}(1 + \tau\delta) \leqslant e^{\delta(t+2\tau)}(|\zeta(0)| + |\zeta(\tau)| + (t+2\tau)B).$$

Le lemme se trouve démontré par suite de l'arbitraire laissé sur t.

L'équation aux différences (2.48) entraîne

$$\eta^\tau(t+2\tau) = \eta^\tau(t) + 2\tau\sigma^\tau(t+\tau)\,\eta^\tau(t+\tau) + 2\tau^{r+1}\xi_3^\tau(t+\tau).$$

Les estimations (2.44), (2.49) aidant, on obtient

$$|\eta^\tau(t+\tau)| \leqslant |\eta^\tau(t)| + |\eta^\tau(t+\tau)|\,2\tau c_{14} + 2\tau^{r+1}c_{18}.$$

D'où, par suite du lemme 2.5 et de (2.53),

$$|\eta^\tau(t)| \leqslant \tau^r e^{2c_{14}t}(c_{20} + 2tc_{18}) \qquad \forall\, t \in \bar{\omega}_\tau.$$

On pose $t = 1$ et on aboutit à l'estimation voulue (2.42), ce qui achève la démonstration du théorème 2.4.

L'estimation (2.54) du lemme 2.5 n'est pas optimale encore que suffisamment simple. Quitte à compliquer la démonstration, on diminue de moitié l'exposant (voir [65]).

Soit de nouveau le système (2.34). On le ramène pour chaque l à deux équations scalaires indépendantes. On fait la somme et la différence des équations (2.34) :

$$z_l' - \frac{\partial f}{\partial u}(t, u)\, z_l = a_l, \qquad t \in [0, 1], \tag{2.55}$$

$$y_l' + \frac{\partial f}{\partial u}(t, u)\, y_l = b_l, \qquad t \in [0, 1], \tag{2.56}$$

où $z_l = v_l + w_l$, $y_l = v_l - w_l$ sont deux nouvelles fonctions inconnues et a_l, b_l deux fonctions connues définies par u, $v_1, \ldots, v_{l-1}$, $w_1, \ldots, w_{l-1}$. Les solutions de ces équations s'écrivent

$$\begin{aligned} z_l(t) &= e^{g(t)}\left(z_l(0) + \int_0^t a_l(x)\, e^{-g(x)}dx\right), \\ y_l(t) &= e^{-g(t)}\left(y_l(0) + \int_0^t b_l(x)\, e^{g(x)}dx\right), \end{aligned} \tag{2.57}$$

avec $g(x) = \int_0^x \frac{\partial f(u(t), t)}{\partial u}\, dt$.

On se place dans le cas où $\partial f/\partial u$ est négative sur l'intervalle d'intégration, ce qui est particulièrement favorable pour la stabilité du problème initial (2.1), (2.2). En effet, les perturbations des données initiales et du second membre s'amortissent de façon exponentielle avec la croissance de t. D'autre part, on déduit de (2.57) que z_l augmente moins vite que a_l et que y_l l'est plus rapidement que b_l. Il y a plus. La fonction y_l a tendance à croître exponentiellement

même pour b_l bornée. Comme $v_l = (z_l + y_l)/2$, $w_l = (z_l - y_l)/2$, les deux fonctions accusent une croissance rapide avec l'augmentation de t. Elles n'ont pas d'intérêt propre du moment que dans l'extrapolation on les élimine des développements (2.39), (2.40). Or le reste η^τ dépend sensiblement de la grandeur de leurs dérivées qui croissent vite quand t augmente.

On rappelle que la somme $z_l = v_l + w_l$ croît en général plus lentement que chacun des termes v_l, w_l pris séparément. Il y a donc intérêt à considérer non (2.39), (2.40), mais un développement qui en est la somme. Comme v_l et w_l ne sont pas définies aux mêmes points du réseau, on recourt à l'interpolation.

La formule de lissage correspondante est (voir [91])

$$\tilde{u}^\tau(1) = \frac{1}{2} u^\tau(1) + \frac{1}{4}[u^\tau(1+\tau) + u^\tau(1-\tau)] = u^\tau_{II}(1). \quad (2.58)$$

Ici $\tilde{u}^\tau(1)$ est une nouvelle valeur « lissée » à l'extrémité de l'intervalle d'intégration. Elle admet un développement de la forme (2.39), (2.40).

Théorème 2.6. *On est dans les hypothèses du théorème 2.4. La fonction « lissée » $\tilde{u}^\tau(1)$ admet le développement*

$$\tilde{u}^\tau(1) = u(1) + \sum_{l=1}^{s} \tau^{2l} g_l + O(\tau^r), \quad (2.59)$$

où $s = [(r-1)/2]$ et les constantes g_l sont indépendantes de τ.

Démonstration. Soit, pour fixer les idées, $M = 1/\tau$ pair. Selon le théorème 2.4, on a pour $t = 1 \pm \tau$

$$u^\tau(t) = u(t) + \sum_{l=1}^{s} \tau^{2l} v_l(t) + \eta^\tau(t).$$

On additionne avec les poids 1/2 et on utilise le lemme 1.1, § 7.1, il vient

$$\frac{1}{2}[u^\tau(1+\tau) + u^\tau(1-\tau)] = \sum_{l=0}^{s} \frac{\tau^{2l}}{(2l)!} u^{(2l)}(1) + \sum_{l=1}^{s} \tau^{2l} \sum_{j=0}^{s-l} \frac{\tau^{2j}}{(2j)!} v_l^{(2j)}(1) + O(\tau^r).$$

On ajoute le développement (2.40) et on divise par 2:

$$u^\tau_{II}(1) = u(1) + \sum_{l=1}^{s} \tau^{2l} \left(\frac{u^{(2l)}(1)}{2(2l)!} + \frac{w_l(1)}{2} + \sum_{j=0}^{l-1} \frac{v_{l-j}^{(2j)}(1)}{2(2j)!} \right) + O(\tau^r).$$

On désigne les expressions correspondantes par g_i et on aboutit à (2.59), c.q.f.d.

On note que le lissage ne résout pas complètement le problème posé par la croissance des coefficients v_i et w_i. Aussi le schéma (2.30), (2.31), (2.58) doublé d'un procédé d'extrapolation à la limite fonctionne au mieux pour un intervalle d'intégration peu important. Supposons qu'on cherche la solution du problème (2.1), (2.2) sur le segment $[0, T]$ grand. On partage naturellement $[0, T]$ en plusieurs segments partiels:

$$[0, T] = [0, x_1] \cup [x_1, x_2] \cup \ldots \cup [x_{p-1}, x_p],$$

où $0 < x_1 < \ldots < x_p = T$. Le problème proposé se décompose en une suite de problèmes

$$\begin{cases} u_1' = f(t, u_1) \quad \text{sur} \quad [0, x_1], \\ u_1(0) = u_0, \end{cases} \tag{2.60}$$

$$\begin{cases} u_i' = f(t, u_i) \quad \text{sur} \quad [x_{i-1}, x_i], \\ u_i(x_{i-1}) = u_{i-1}(x_{i-1}), \\ i = 2, 3, \ldots, p. \end{cases} \tag{2.61}$$

Il est clair que les solutions exactes de ces problèmes coïncident avec u sur les segments correspondants. On cherche chaque solution par le schéma (2.30), (2.31), (2.58) et un procédé d'extrapolation. On initialise (2.61) avec la valeur approchée $u_{i-1}^{\tau}(x_{i-1})$ résultant du problème précédent.

Bulirsch et Stoer ont proposé de choisir la longueur des intervalles d'intégration en cours de calcul, ce choix étant fonction de la précision voulue et de l'ordre de l'extrapolation. Les algorithmes de ces auteurs utilisent l'extrapolation rationnelle. Ils sont efficaces dans le cas de problèmes à données régulières lorsqu'on demande une solution approchée hautement précise. Pour plus de détails, voir [75], [76], [117].

Remarque. La généralisation a lieu pour diverses conditions aux limites. Ainsi, on cherche dans [99], [130] la solution de l'équation scalaire $y' = f(t, y)$ sur l'intervalle (a, b), qui vérifie la condition $g(y(a), y(b)) = 0$, relation non linéaire entre deux valeurs de la fonction inconnue. Avec le schéma de Crank-Nicholson, on obtient un développement suivant les puissances paires du paramètre de discrétisation qui permet d'utiliser l'algorithme de raffinement décrit au § 2.1.

Dans [97], on établit un développement sous forme vectorielle suivant les puissances paires pour résoudre numériquement $\mathbf{y}' = A\mathbf{y} + \mathbf{g}$ sur (a, b). Ici $\mathbf{y}$ et $\mathbf{g}$ sont des vecteurs de n composantes

et A une matrice carrée d'ordre n. Les composantes des vecteurs et les éléments de la matrice sont des fonctions de l'argument t. On n'introduit pas de condition initiale usuelle, et on garantit l'unicité par la contrainte

$$\sum_{i=1}^{N} B_i \mathbf{y}(t_i) = \boldsymbol{\beta},$$

$\boldsymbol{\beta}$ étant un vecteur connu de n composantes et $\{B_i\}$ une famille de matrices à n lignes et n colonnes. On discrétise de façon que le réseau renferme les nœuds t_i. L'approximation de l'équation différentielle s'effectue par le schéma de Crank-Nicholson.

2.3. Méthode de décomposition pour un système d'équations

S'agissant du problème de Cauchy pour un système d'équations différentielles linéaires, il y a parfois intérêt à réduire le problème initial à une suite de problèmes de Cauchy plus simples. C'est le cas des systèmes dont la matrice des coefficients est représentable sous forme de somme de matrices plus simples douées de propriétés déterminées. La réduction est alors réalisée par la méthode de décomposition (voir [43], [112], [141]). Cette technique s'avère efficace pour les problèmes élémentaires si l'on utilise plusieurs réseaux successifs. On obtient finalement une solution améliorée.

Soit le système d'équations différentielles linéaires

$$\frac{d\mathbf{u}}{dt} + A\mathbf{u} = \mathbf{f}, \qquad t \in (0, 1), \tag{3.1}$$

avec la condition initiale

$$\mathbf{u}(0) = \mathbf{u}_0. \tag{3.2}$$

La matrice $A(t)$ est carrée d'ordre m, $\mathbf{f}(t)$, $\mathbf{u}(t)$ et $\mathbf{u}_0$ sont des vecteurs de m composantes. On suppose que les éléments de A et les composantes de $\mathbf{f}$ sont de classe $C^r[0, 1]$, auquel cas (voir [77]) il y a existence et unicité, et le vecteur $\mathbf{u}$ a ses composantes dans $C^{r+1}[0,1]$.

On admet que A se met sous forme de somme de matrices à m lignes et m colonnes dont les éléments présentent la même régularité:

$$A = \sum_{i=1}^{n} A_i, \tag{3.3}$$

les matrices $A_i(t)$ étant définies non négatives à tout instant $t \in [0, 1]$:

$$(A_i(t)\,\mathbf{v}, \mathbf{v}) \geqslant 0 \qquad \forall\, \mathbf{v} \in \mathbf{E}^m. \tag{3.4}$$

Ici $\mathbf{E}^m$ est l'espace vectoriel de dimension m muni du produit scalaire

$$(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{m} v_i w_i.$$

où

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ \cdot \\ v_m \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ \cdot \\ \cdot \\ \cdot \\ w_m \end{bmatrix},$$

et de la norme

$$\|\mathbf{v}\| = (\mathbf{v}, \mathbf{v})^{1/2}.$$

La résolution numérique du problème (3.1), (3.2) se fait par le schéma décomposé implicite (voir [112])

$$(I + \tau A_1(t))\mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right) = \mathbf{u}^\tau(t - \tau) + \tau\mathbf{f}(t),$$

$$(I + \tau A_2(t))\mathbf{u}^\tau\left(t - \tau\frac{n-2}{n}\right) = \mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right),$$

$$\cdots\cdots\cdots\cdots\cdots\cdots \qquad (3.5)$$

$$(I + \tau A_n(t))\,\mathbf{u}^\tau(t) = \mathbf{u}^\tau\left((t - \tau\frac{1}{n}\right),$$

$$t \in \omega_\tau,$$

$$\mathbf{u}^\tau(0) = \mathbf{u}_0. \qquad (3.6)$$

I étant une matrice unité à m lignes et m colonnes.

On note qu'on résout à chaque pas les systèmes d'équations algébriques linéaires de matrice $I + \tau A_i(t)$, si bien qu'on décompose A de façon que la résolution des systèmes soit facile.

On démontre la stabilité du système (3.5), (3.6).

THÉORÈME 3.1. *On a pour le problème* (3.5), (3.6) *l'estimation à priori*

$$\max_{t \in \bar{\omega}_\tau} \|\mathbf{u}^\tau(t)\| \leqslant \|\mathbf{u}_0\| + \max_{t \in \omega_\tau} \|\mathbf{f}(t)\|. \qquad (3.7)$$

DÉMONSTRATION. On multiplie scalairement chaque équation (3.5) par $\mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right), \ldots, \mathbf{u}^\tau(t)$ respectivement. La première équation donne

$$\left\|\mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right)\right\|^2 + \tau\left(A_1(t)\,\mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right), \mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right)\right) =$$

$$= \left(\mathbf{u}^\tau(t - \tau), \mathbf{u}^\tau\left(t - \tau\frac{n-1}{n}\right)\right) + \tau\left(\mathbf{f}(t), \mathbf{u}^\tau t - \tau\frac{n-1}{n}\right)\right).$$

On utilise la propriété de A_1 d'être définie non négative et l'inégalité de Cauchy-Bouniakovski, il vient

$$\left\| u^\tau\left(t - \tau\frac{n-1}{n}\right)\right\|^2 \leqslant \| u^\tau(t-\tau)\| \left\| u^\tau\left(t - \tau\frac{n-1}{n}\right)\right\| + \\ + \tau \| f(t)\| \left\| u^\tau\left(t - \tau\frac{n-1}{n}\right)\right\|.$$

Cette inégalité entraîne, on le voit aisément, l'estimation

$$\left\| u^\tau\left(t - \tau\frac{n-1}{n}\right)\right\| \leqslant \| u^\tau(t-\tau)\| + \tau \| f(t)\|. \qquad (3.8)$$

On traite de même les autres équations, et on a la chaîne d'inégalités

$$\left\| u^\tau\left(t - \tau\frac{n-2}{n}\right)\right\| \leqslant \left\| u^\tau\left(t - \tau\frac{n-1}{n}\right)\right\|,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\| u^\tau(t)\| \leqslant \left\| u^\tau\left(t - \tau\frac{1}{n}\right)\right\|.$$

Ces inégalités et l'estimation (3.8) donnent

$$\| u^\tau(t)\| \leqslant \| u^\tau(t-\tau)\| + \tau \| f(t)\|.$$

On en déduit par recours à $\| u^\tau(0)\| = \| u_0\|$ résultant de la condition initiale (3.6):

$$\| u^\tau(t)\| \leqslant \| u_0\| + t \max_{t\in\omega_\tau} \| f(t)\|.$$

D'où l'estimation voulue (3.7).

Les pages suivantes de ce paragraphe généralisent les résultats du chapitre premier au cas de vecteurs, si bien que les démonstrations se font en gros suivant les mêmes schémas. Nous insisterons sur les différences capitales entre le cas vectoriel et le cas scalaire.

Soit l'ensemble de problèmes différentiels*

$$\frac{dv_l}{dt} + A\, v_l = -\sum_{s=2}^{\min(l+1,n)} \left(\sum_{i_1<i_2<\ldots<i_s} A_{i_1}\ldots A_{i_s}\right) v_{l-s+1} + \\ + \sum_{i=0}^{l-1} \frac{(-1)^{l-i+1}}{(l-i+1)!}\, v_i^{(l-i+1)} \quad \text{sur} \quad (0,\ 1) \qquad (3.9)$$

---

*) On considère comme étant nulles les sommes dont la limite inférieure est plus grande que la limite supérieure. Le signe $\sum\limits_{i_1<i_2<\ldots<i_s}$ signifie la sommation de tous les produits (distincts) possibles de s facteurs $A_{i_1} A_{i_2} \ldots A_{i_s}$ tels que $1 \leqslant i_1 < i_2 < \ldots < i_s \leqslant n$.

et

$$\mathbf{v}_l(0) = 0^*, \qquad l = 1, \ldots, r. \tag{3.10}$$

Si l'on pose $\mathbf{v}_0 = \mathbf{u}$, on obtient à partir de ce système les fonctions vectorielles successives $\mathbf{v}_l$ dans l'ordre de croissance de l'indice l.

En effet, on suppose définies k fonctions vectorielles $\mathbf{v}_i$, $i = 0, \ldots, k-1$, de composantes $\in C^{r+1-i}[0, 1]$. Le plus grand indice des fonctions figurant au second membre de (3.9) est $k - 1$; aussi la fonction vectorielle $\mathbf{v}_k$ est bien obtenue à partir de (3.9) avec la condition initiale homogène. On note en outre que les composantes du second membre sont $r - k$ fois continûment dérivables sur $[0,1]$. La solution $\mathbf{v}_k$ a donc ses composantes dans $C^{r-k+1}[0, 1]$. Par conséquent, le système (3.9), (3.10) fournit univoquement toutes les $\mathbf{v}_l$, et les composantes de ces fonctions sont de classes $C^{r+1-l}[0, 1]$.

On montre que u^τ du schéma (3.5) admet la représentation

$$\mathbf{u}^\tau = \sum_{j=0}^{r-1} \tau^j \mathbf{v}_j + \tau^r \boldsymbol{\eta}^\tau \quad \text{sur} \quad \bar{\omega}_\tau, \tag{3.11}$$

avec la fonction vectorielle η^τ définie aux nœuds de $\bar{\omega}_\tau$ telle que

$$\max_{t \in \bar{\omega}_\tau} \| \boldsymbol{\eta}^\tau(t) \| \leqslant c_1. \tag{3.12}$$

Les fonctions $\mathbf{u}^\tau$ et $\mathbf{v}_j$ étant définies, l'égalité (3.11) donne la définition suivante de $\boldsymbol{\eta}^\tau$:

$$\boldsymbol{\eta}^\tau = \mathbf{u}^\tau - \sum_{j=0}^{r-1} \tau^j \mathbf{v}_j \quad \text{sur} \quad \bar{\omega}_\tau. \tag{3.13}$$

Il reste à prouver l'estimation (3.12). On récrit le système (3.5) en en éliminant les valeurs intermédiaires de $\mathbf{u}^\tau$:

$$(I + \tau A_1(t)) \ldots (I + \tau A_n(t)) \mathbf{u}^\tau(t) = \mathbf{u}^\tau(t - \tau) + \\ + \tau \mathbf{f}(t), \qquad t \in \bar{\omega}_\tau. \tag{3.14}$$

On y porte le développement (3.11), il vient

$$(I + \tau A_1(t)) \ldots (I + \tau A_n)(t)) \left(\sum_{j=0}^{r-1} \tau^j \mathbf{v}_j(t) + \tau^r \boldsymbol{\eta}^\tau(t) \right) = \\ = \sum_{j=0}^{r-1} \tau^j \mathbf{v}_j(t - \tau) + \tau^r \boldsymbol{\eta}^\tau(t - \tau) + \tau \mathbf{f}(t). \tag{3.15}$$

---

*) Ici 0 est l'élément zéro de l'espace E^m.

On développe chaque $\mathbf{v}_j$ du second membre en série de Taylor avec un reste qui en exprime la régularité:

$$\mathbf{v}_j(t-\tau) = \sum_{i=0}^{r-j} \frac{(-\tau)^i}{i!} \mathbf{v}_j^{(i)}(t) + \tau^{r-j+1} \boldsymbol{\sigma}_j^\tau(t),$$

où

$$\| \boldsymbol{\sigma}_j^\tau(t) \| \leqslant \frac{\sqrt{m}}{(r-j+1)!} \max_{1 \leqslant k \leqslant m} \max_{[0,1]} \left| v_{j,i}^{(r-j+1)} \right| \; \forall t \in \omega_\tau, \tag{3.16}$$

$v_{j,i}$ étant la i-ième composante de $\mathbf{v}_j$. Avec la formule pour $\mathbf{v}_j$ $(t - \tau)$, on récrit (3.15) en omettant la variable indépendante t dans les cas où cela n'entraîne aucune ambiguïté. On a

$$(I + \tau A_1) \ldots (I + \tau A_n) \left(\sum_{j=0}^{r-1} \tau^j \mathbf{v}_j + \tau^r \boldsymbol{\eta}^\tau \right) =$$

$$= \sum_{j=0}^{r-1} \tau^j \sum_{i=0}^{r-j} \frac{(-\tau)^i}{i!} \mathbf{v}_j^{(i)} + \tau^r \boldsymbol{\eta}^\tau (t-\tau) + \tau \mathbf{f} + \tau^{r+1} \sum_{j=0}^{r-1} \boldsymbol{\sigma}_j^\tau.$$

On multiplie les termes du premier membre et on identifie les termes contenant les mêmes puissances de τ:

$$\sum_{s=1}^{n} \tau^s \sum_{i_1 < i_2 < \ldots < i_s} A_{i_1} A_{i_2} \ldots A_{i_s} \sum_{j=0}^{r-1} \tau^j \mathbf{v}_j +$$

$$+ \tau^r (I + \tau A_1) \ldots (I + \tau A_n) \boldsymbol{\eta}^\tau =$$

$$= \sum_{j=0}^{r} \tau^j \sum_{i=0}^{j-1} \frac{(-1)^{j-i}}{(j-i)!} \mathbf{v}_{(i)}^{(j-i)} + \tau^r \boldsymbol{\eta}^\tau (t-\tau) + \tau \mathbf{f} + \tau^{r+1} \sum_{j=0}^{r-1} \boldsymbol{\sigma}_j^\tau.$$

On réduit aisément cette relation à la forme

$$\sum_{j=1}^{r+n-1} \tau^j \sum_{s=1}^{\min(j,n)} \sum_{i_1 < i_2 < \ldots < i_s} A_{i_1} A_{i_2} \ldots A_{i_s} \mathbf{v}_{j-s} +$$

$$+ \tau^r (I + \tau A_1) \ldots (I + \tau A_n) \boldsymbol{\eta}^\tau = \sum_{j=0}^{r} \tau^j \sum_{i=0}^{j-1} \frac{(-1)^{j-i}}{(j-i)!} \mathbf{v}_i^{(j-i)} +$$

$$+ \tau^r \boldsymbol{\eta}^\tau (t-\tau) + \tau \mathbf{f} + \tau^{r+1} \sum_{j=0}^{r-1} \boldsymbol{\sigma}_j^\tau. \tag{3.17}$$

Le procédé de recherche de $\mathbf{v}_l$ (les égalités (3.9)) nous fait conclure facilement à la validité des équations

$$\frac{d\mathbf{v}_0}{dt} + A\,\mathbf{v}_0 = \mathbf{f} \quad \text{sur} \quad (0,\ 1),$$

car $\mathbf{v}_0 = \mathbf{u}$ et

$$\sum_{s=1}^{\min(l+1,n)} \sum_{i_1<i_2<\ldots<i_s} A_{i_1} A_{i_2} \ldots A_{i_s} \mathbf{v}_{l-s+1} =$$

$$= \sum_{i=0}^{l} \frac{(-1)^{l-i+1}}{(l-i+1)!} \mathbf{v}_i^{(l-i+1)} \quad \text{sur} \quad (0,\ 1),\ l = 1,\ \ldots,\ r-1.$$

On note qu'il s'agit en fait d'une autre écriture des équations (3.9). On utilise les relations obtenues pour simplifier les deux membres de (3.17):

$$\sum_{j=r+1}^{r+n-1} \tau^j \sum_{s=1}^{\min(j,n)} \sum_{i_1<i_2<\ldots<i_s} A_{i_1} A_{i_2} \ldots A_{i_s} \mathbf{v}_{j-s} +$$

$$+ \tau^r (I + \tau A_1) \ldots (I + \tau A_n)\, \boldsymbol{\eta}^\tau = \tau^r \boldsymbol{\eta}^\tau (t - \tau) + \tau^{r+1} \sum_{j=0}^{r-1} \boldsymbol{\sigma}_j^\tau;$$

on reporte à gauche le premier groupe de termes du second membre:

$$(I + \tau A_1) \ldots (I + \tau A_n)\, \boldsymbol{\eta}^\tau = \boldsymbol{\eta}^\tau (t - \tau) + \tau \boldsymbol{\rho}^\tau, \qquad t \in \omega_\tau, \quad (3.18)$$

où la fonction vectorielle discrète $\boldsymbol{\rho}^\tau$ est uniformément bornée sur ω_τ:

$$\max_{t \in \omega_\tau} \| \boldsymbol{\rho}^\tau (t) \| \leqslant c_2. \qquad (3.19)$$

La dernière inégalité tient à ce que $\boldsymbol{\rho}^\tau$ est la somme de termes de deux types:

$$\boldsymbol{\rho}^\tau = \sum_{j=0}^{r-1} \boldsymbol{\sigma}_j^\tau - \sum_{j=r+1}^{r+n-1} \tau^{j-r-1} \sum_{s=1}^{\min(j,n)} \sum_{i_1<i_2<\ldots<i_s} A_{i_1} A_{i_2} \ldots A_{i_s} \mathbf{v}_{j-s}. \qquad (3.20)$$

L'estimation (3.16) et la régularité suffisante des composantes des $\mathbf{v}_j$ déterminent la borne uniforme sur ω_τ de la première somme. La seconde somme a tous ses termes bornés parce que les éléments des matrices A_i et les composantes des vecteurs $\mathbf{v}_{j-s}$ présentent la

propriété d'être uniformément bornés. En effet, ces éléments et ces composantes sont au moins continus, et ils atteignent donc leur maximum et leur minimum sur [0, 1]. Les termes du second groupe sont au plus $(n-1)2^n$. On évalue chaque terme. Soit $\|\tau^{j-r-1} A_{i_1} \ldots \ldots A_{i_s} \mathbf{v}_{j-s}\|$. Comme $j \geqslant r+1$ et $\tau < 1$, cette expression est au plus égale à $\|A_{i_1} \ldots A_{i_s} \mathbf{v}_{j-s}\|$. On choisit la composante maximum du vecteur $\mathbf{v}_{j-s}$ (il est connu que les composantes de ce vecteur sont uniformément bornées):

$$c_3 = \max_{1 \leqslant k \leqslant m} \max_{[0,1]} |v_{j-s,k}|,$$

avec $v_{j-s,k}$ la k-ième composante de $\mathbf{v}_{j-s}$. Il est clair que $\|\mathbf{v}_{j-s}\| \leqslant \leqslant c_3 \sqrt{m}$. On prend l'élément de plus grand module de la matrice A_i:

$$c_4 = \max_{1 \leqslant i \leqslant n} \max_{[0,1]} \max_{\substack{1 \leqslant k \leqslant m \\ 1 \leqslant l \leqslant m}} |A_{i,k,l}|,$$

avec $A_{i,k,l}$ l'élément en ligne k et en colonne l. On utilise pour la norme euclidienne de A_i l'inégalité fort simple (voir par exemple [119])

$$\|A_i\| \leqslant mc_4,$$

et on a finalement

$$\|A_{i_1} \ldots A_{i_s} \mathbf{v}_{j-s}\| \leqslant \|A_{i_1}\| \cdots \|A_{i_s}\| \, \|\mathbf{v}_{j-s}\| \leqslant m^s c_4^s c_3.$$

Ainsi, on a évalué un terme arbitraire de la somme (3.20), d'où la validité de (3.19).

Avec la définition (3.13), on construit la condition initiale pour l'équation (3.18) au point $t = 0$. Cette condition s'écrit compte tenu de (3.2), (3.6), (3.10),

$$\boldsymbol{\eta}^\tau(0) = \tau^{-r}\left(\mathbf{u}^\tau(0) - \sum_{j=0}^{r-1} \tau^j \mathbf{v}_j(0)\right) = \tau^{-r}(\mathbf{u}(0) - \mathbf{u}(0)) = 0.$$

On a donc construit pour $\boldsymbol{\eta}^\tau$ un problème aux différences qui vérifie l'estimation à priori (3.7). L'estimation (3.7) implique

$$\max_{t \in \bar{\omega}_\tau} \|\boldsymbol{\eta}^\tau(t)\| \leqslant \max_{t \in \omega_\tau} \|\boldsymbol{\rho}^\tau(t)\|.$$

Etant donné (3.19), on est conduit à (3.12), et $c_1 = c_2$.

On énonce ce résultat sous forme de

THÉORÈME 3.2. *On suppose que la matrice A du problème* (3.1), (3.2) *s'écrit sous forme de somme* (3.3) *de n termes vérifiant la condition* (3.4) *et que les éléments des matrices A_i et les composantes*

du vecteur $\mathbf{f}$ *appartiennent à* C^r $[0, 1]$. *La solution* $\mathbf{u}^\tau$ *du schéma décomposé* (3.5), (3.6) *admet le développement*

$$\mathbf{u}^\tau = \mathbf{u} + \sum_{j=1}^{r-1} \tau^j \mathbf{v}_j + \tau^r \boldsymbol{\eta}^\tau \quad \text{sur} \quad \bar{\omega}_\tau,$$

où $\mathbf{v}_j$ *est une fonction vectorielle de composantes* $\in C^{r+1-j}$ $[0, 1]$ *indépendantes de* τ *et* $\boldsymbol{\eta}^\tau$ *une fonction vectorielle discrète bornée :*

$$\max_{t \in \bar{\omega}_\tau} \| \boldsymbol{\eta}^\tau (t) \| \leqslant c_1. \tag{3.21}$$

On se base sur ce développement pour justifier la méthode d'extrapolation du § 1.3 à la différence qu'on opère avec les vecteurs et non avec les scalaires. On suppose qu'on est dans les hypothèses du théorème 3.2 et on construit pour $0 < N_1 < \ldots < N_r$ entiers fixés les réseaux $\bar{\omega}_{\tau_k}$ de pas $\tau_k = 1/(N_k M)$, où M est un entier naturel qui croît indéfiniment. On pose pour chaque réseau $\bar{\omega}_{\tau_k}$ le problème approché (3.5), (3.6). Leurs solutions $\mathbf{u}^{\tau_k}$ sont définies sur $\bar{\omega}_\tau$ de pas $\tau = 1/M$. Le système d'équations

$$\begin{gathered} \sum_{k=1}^{r} \gamma_k = 1, \\ \sum_{k=1}^{r} \gamma_k \tau_k^j = 0, \qquad j = 1, \ldots, r-1, \end{gathered} \tag{3.22}$$

a son déterminant différent de 0, si bien qu'il existe une solution unique $\gamma_1, \ldots, \gamma_r$. On forme la combinaison linéaire

$$\mathbf{U}^H (t) = \sum_{k=1}^{r} \gamma_k \mathbf{u}^{\tau_k} (t), \qquad t \in \bar{\omega}_\tau, \tag{3.23}$$

et on démontre sa précision plus grande pour $\tau \to 0$.

THÉORÈME 3.3. *On suppose que le problème* (3.1), (3.2) *satisfait aux conditions* (3.3), (3.4) *et que les éléments des matrices* A_i *et les composantes du vecteur* $\mathbf{f}$ *sont dans* C^r $[0, 1]$, *avec* r *entier naturel. La solution corrigée* (3.23) *avec les poids* γ_k *vérifiant le système* (3.22) *admet l'estimation*

$$\max_{t \in \bar{\omega}_\tau} \|\mathbf{U}^H (t) - \mathbf{u}(t) \| \leqslant c_3 \tau^m. \tag{3.24}$$

DÉMONSTRATION. On est dans les hypothèses du théorème 3.2, si bien qu'on a en chaque nœud du réseau $\bar{\omega}_\tau$

$$\mathbf{u}^{\tau_k} = \mathbf{u} + \sum_{j=1}^{r-1} \tau^j \mathbf{v}_j + \tau_k^r \boldsymbol{\eta}^{\tau_k}, \qquad k = 1, \ldots, r.$$

On additionne ces développements munis des poids γ_k et on prend en considération la propriété (3.22) des coefficients γ_k, il vient

$$\mathbf{U}^H = \mathbf{u} + \sum_{k=1}^{r} \gamma_k \tau_k^r \boldsymbol{\eta}^{\tau_k} \qquad \text{sur } \bar{\omega}_\tau. \tag{3.25}$$

On évalue $|\gamma_k|$ à l'aide du lemme 2.3, § 7.2. On pose

$$c_6 = \min_{1 \leq k \leq r-1} \frac{N_{k+1}}{N_k} - 1,$$

auquel cas

$$|\gamma_k| \leq \left(\frac{1+c_6}{c_6}\right)^r, \qquad k = 1, \ldots, r.$$

Cette inégalité et l'estimation (3.21) permettent de déduire de (3.25)

$$\| \mathbf{U}^H(t) - \mathbf{u}(t) \| \leq c_1 \left(\frac{1+c_6}{c_6}\right)^r \sum_{k=1}^{r} \tau_k^r, \qquad t \in \bar{\omega}_\tau.$$

Mais $\tau_k \leq \tau$, $k = 1, \ldots, r$, si bien qu'avec la notation

$$c_5 = r c_1 \left(\frac{1+c_6}{c_6}\right)^r,$$

on aboutit à (3.24), c.q.f.d.

Explicitons ces résultats sur un exemple numérique. Soit le problème (3.1), (3.2) de A symétrique:

9,6045364	−0,2154636	−0,1974636	−0,7826544	−1,7593524
	0,9645364	−0,0174636	−0,0626544	−0,1393524
éléments		0,1005364	0,0093456	0,0226476
symétriques			0,0761824	0,1553904
des éléments				0,3652084
supérieurs				

Ses valeurs propres sont égales à 10, 1, 0,1, 0,01, 0,001. La matrice A étant symétrique et à valeurs propres positives est définie non négative (voir [136]). On décompose A en deux matrices triangulaires A_1 et A_2, la matrice triangulaire inférieure A_1 étant formée d'éléments subdiagonaux correspondants et de valeurs divisées par 2 des éléments diagonaux et la matrice triangulaire supérieure A_2

comportant les éléments surdiagonaux correspondants et les valeurs divisées par 2 des éléments diagonaux. Il est clair que $A_1 = A_2^T$ et $A = A_1 + A_2$. On démontre que A_1 et A_2 sont définies non négatives:

$$(A_i \mathbf{u}, \mathbf{u}) = \frac{1}{2}(A_i \mathbf{u}, \mathbf{u}) + \frac{1}{2}(\mathbf{u}, A_i^T \mathbf{u}) = \frac{1}{2}((A_i + A_i^T)\mathbf{u}, \mathbf{u}) =$$

$$= \frac{1}{2}(A\mathbf{u}, \mathbf{u}) \geqslant 0 \qquad \forall\, \mathbf{u} \in \mathbf{E}^5, \qquad i = 1,2.$$

Le vecteur valeurs initiales a été pris égal à

$$(0{,}68,\ 0{,}68,\ 0{,}68,\ -0{,}28,\ -1{,}88)^T.$$

La solution exacte du problème (3.1), (3.2) avec second membre nul est la fonction vectorielle

$$\mathbf{u}(t) = \begin{bmatrix} 0{,}98 & -0{,}02 & -0{,}02 & -0{,}08 & -0{,}18 \\ -0{,}02 & 0{,}98 & -0{,}02 & -0{,}08 & -0{,}18 \\ -0{,}02 & -0{,}02 & 0{,}98 & -0{,}08 & -0{,}18 \\ -0{,}08 & -0{,}08 & -0{,}08 & 0{,}68 & -0{,}72 \\ -0{,}18 & -0{,}18 & -0{,}18 & -0{,}72 & -0{,}62 \end{bmatrix} \begin{bmatrix} e^{-10t} \\ e^{-t} \\ e^{-0,1t} \\ e^{-0,01t} \\ e^{-0,001t} \end{bmatrix}.$$

Le tableau 2.1 donne les erreurs sur les solutions discrètes du problème (3.5), (3.6) et sur les solutions extrapolées (3.23).

Tableau 2.1

no de la composante	Erreur maximum sur la solution discrète du problème (3.5), (3.6)			Erreur maximum dans l'extrapolation (3.23)	
	$\tau = \frac{1}{80}$	$\tau = \frac{1}{160}$	$\tau = \frac{1}{320}$	$\tau_1 = \frac{1}{80}$, $\tau_2 = \frac{1}{160}$	$\tau_1 = \frac{1}{80}$, $\tau_2 = \frac{1}{160}$, $\tau_3 = \frac{1}{320}$
1	$1{,}55 \cdot 10^{-2}$	$7{,}87 \cdot 10^{-3}$	$3{,}97 \cdot 10^{-3}$	$2{,}35 \cdot 10^{-4}$	$1{,}83 \cdot 10^{-6}$
2	$2{,}85 \cdot 10^{-3}$	$1{,}44 \cdot 10^{-3}$	$7{,}21 \cdot 10^{-4}$	$2{,}21 \cdot 10^{-5}$	$1{,}86 \cdot 10^{-7}$
3	$8{,}96 \cdot 10^{-4}$	$4{,}54 \cdot 10^{-4}$	$2{,}29 \cdot 10^{-4}$	$1{,}28 \cdot 10^{-5}$	$1{,}38 \cdot 10^{-7}$
4	$3{,}42 \cdot 10^{-3}$	$1{,}73 \cdot 10^{-3}$	$8{,}72 \cdot 10^{-4}$	$4{,}97 \cdot 10^{-5}$	$5{,}40 \cdot 10^{-7}$
5	$7{,}74 \cdot 10^{-3}$	$3{,}93 \cdot 10^{-3}$	$1{,}98 \cdot 10^{-3}$	$1{,}18 \cdot 10^{-4}$	$1{,}29 \cdot 10^{-6}$

2.4. Equations avec singularités

On s'est borné jusqu'à présent au cas d'équations différentielles ayant des solutions régulières. Dans ce paragraphe, on traitera le problème de rechercher les solutions des équations différentielles avec singularités.

On démontrera que ce cas se prête également à l'extrapolation de Richardson à partir des solutions approchées relativement peu précises associées à une succession de réseaux. La méthode de Richardson sera basée sur le développement de la solution approchée suivant les puissances fractionnaires du pas de discrétisation. On examinera un exemple concret pour illustrer les procédés de recherche de ses premiers termes qu'on développe dans [87] et [116]. Si les équations différentielles possèdent des singularités d'autres types (disons, des singularités logarithmiques), la méthode d'extrapolation s'avère opérante dans certains cas.

Soit le problème

$$u' = \sqrt{t}, \qquad t \in (0, 1), \tag{4.1}$$

$$u(0) = 0, \tag{4.2}$$

dont la solution est la fonction

$$u = \frac{2}{3} t^{3/2}. \tag{4.3}$$

Bien que la régularité nécessaire de la solution sur le segment [0, 1] fasse défaut, on procède par le schéma de Crank-Nicholson

$$u_{\bar{t}}^{\tau} = \sqrt{\bar{t}}, \qquad t \in \bar{\omega}_{\tau}, \tag{4.4}$$

$$u^{\tau}(0) = 0. \tag{4.5}$$

Quel est le comportement sur $\bar{\omega}_{\tau}$ de l'erreur $\psi^{\tau} = u^{\tau} - u$? On porte $u^{\tau} = u + \psi^{\tau}$ dans (4.4), il vient

$$u_{\bar{t}} + \psi_{\bar{t}}^{\tau} = \bar{t}^{1/2}, \qquad t \in \bar{\omega}_{\tau}. \tag{4.6}$$

La fonction u admet des dérivées de tous ordres sur tout intervalle ne contenant pas le point $t = 0$, si bien qu'on a en tous les points de $\bar{\omega}_{\tau}$, sauf le premier:

$$u_{\bar{t}}(t) = u'(\bar{t}) + \frac{\tau^2}{24} u'''(\bar{t}) + \frac{\tau^4}{16 \cdot 120} u^{5}(\xi) =$$

$$= \bar{t}^{1/2} - \frac{\tau^2}{96} \bar{t}^{-3/2} - \frac{\tau^4}{2048} \xi^{-7/2},$$

où $\xi(t) \in [t - \tau/2,\ t + \tau/2]$. Etant donné ce développement, l'équation (4.6) se récrit

$$\psi_t^\tau = +\frac{\tau^2}{96}\, t^{-3/2} + \frac{\tau^4}{2048}\, \xi^{-7/2}. \tag{4.7}$$

Il est évident que le terme en τ^2 est prédominant pour tous les points de $\bar{\omega}_\tau$, sauf le premier. On procède comme plus haut, et on dégage la partie principale de ψ^τ moyennant la solution de l'équation différentielle

$$v' = \frac{\tau^2}{96}\, t^{-3/2}, \qquad t \in (\tau,\ 1). \tag{4.8}$$

La solution générale de (4.8) est la fonction

$$v = -\frac{\tau^2}{48}\, t^{-1/2} + d, \tag{4.9}$$

avec d une constante indépendante de t. Il est clair que la condition initiale $v(0) = 0$ n'est remplie pour aucune d. On établit donc la valeur de d à l'aide de

$$u_t^\tau\left(\frac{\tau}{2}\right) = \sqrt{\frac{\tau}{2}}.$$

Comme $u^\tau(0) = 0$, on a $u^\tau(\tau) = \tau^{3/2}/\sqrt{2}$, d'où

$$u(\tau) + \psi^\tau(\tau) = \tau^{3/2}/\sqrt{2}.$$

Si l'on veut que v décrive le comportement de la partie principale de l'erreur, il faut donc que

$$v(\tau) = \psi^\tau(\tau) = \frac{3\sqrt{2} - 4}{6}\, \tau^{3/2},$$

d'où

$$d = \frac{24\sqrt{2} - 31}{48}\, \tau^{3/2}.$$

Par conséquent, la partie principale de l'erreur est définie sur ω_τ par la formule

$$v(t) = \frac{24\sqrt{2} - 31}{48}\, \tau^{3/2} - \frac{\tau^2}{48}\, t^{-1/2}.$$

Les résultats de l'expérience numérique correspondante sont donnés sous forme de relations entre les quantités

$$\eta_1(M) = |\psi^\tau(1)|, \tag{4.10}$$

$$\eta_2(M) = \psi^\tau(1) - \frac{24\sqrt{2} - 31}{48}\, \tau^{3/2} \tag{4.11}$$

et $M = 1/\tau$ entier. La fig. 2.2 visualise ces relations en coordonnées logarithmiques. La coïncidence des pentes théoriques et réelles illustre bien l'ordre de grandeur établi antérieurement de η_1 et η_2.

On note que s'agissant du développement de la solution suivant les puissances fractionnaires, on améliore la précision par l'extrapolation de Richardson. En effet, on fixe l entiers naturels $N_1 < \ldots$ $\ldots < N_l$ et on suppose que les solutions approchées ont en t pour tout M

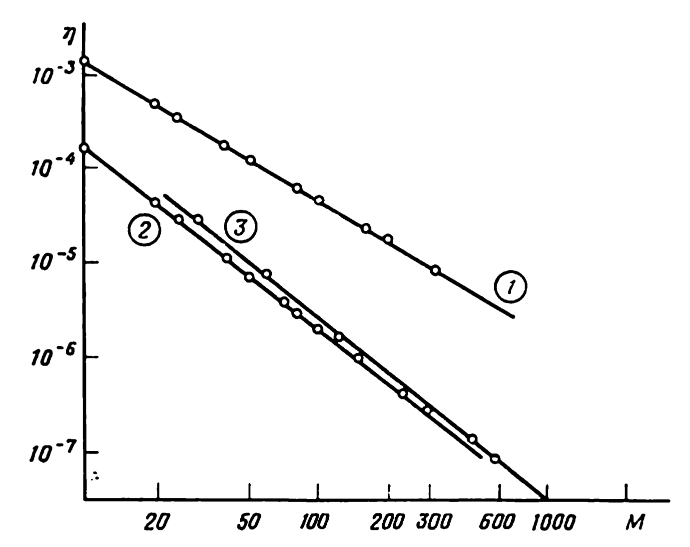


Fig. 2.2. Erreurs sur les solutions approchées du problème (4.1), (4.2) pour $t = 1$.
1 — η_1 de (4.10); *2* — η_2 de (4.11); *3* — η_3 de (4.17).

entier naturel l développements suivant le paramètre de discrétisation $\tau_k = \tau/N_k$, où $\tau = 1/M$:

$$u^{\tau_k}(t) = u(t) + \sum_{i=1}^{l-1} \tau_k^{\alpha_i} v_i(t) + \tau_k^{\alpha_l} \eta^{\tau_k}(t). \tag{4.12}$$

$$k = 1, \ldots, l.$$

Les fonctions u, v_i et les exposants α_i, α_l ne dépendent pas de τ, et $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_l$. Le reste vérifie de plus l'estimation

$$\left|\eta^{\tau_k}(t)\right| \leqslant c_1. \tag{4.13}$$

On forme la combinaison linéaire

$$U^H(t) = \sum_{k=1}^{l} \gamma_k u^{\tau_k}(t) \tag{4.14}$$

avec les poids vérifiant le système

$$\sum_{k=1}^{l} \gamma_k = 1,$$
$$\sum_{k=1}^{l} \frac{\gamma_k}{N_k^{\alpha_i}} = 0, \qquad i = 1, \ldots, l-1. \tag{4.15}$$

Démontrer la possibilité du système (4.15) exigerait une information supplémentaire abondante, si bien que nous ferons tout simplement l'hypothèse qu'il possède pour N_k choisis une solution $\gamma_1, \ldots, \gamma_l$. Il y a lieu de dire que le système et la solution sont indépendants du paramètre M. Le système (4.15) entraîne les égalités

$$\sum_{k=1}^{l} \gamma_k = 1,$$
$$\sum_{k=1}^{l} \gamma_k \tau_k^{\alpha_i} = 0, \qquad i = 1, \ldots, l-1. \tag{4.16}$$

On a par suite de (4.12)

$$U^H(t) = \sum_{k=1}^{l} \gamma_k u(t) + \sum_{k=1}^{l} \sum_{i=1}^{l-1} \gamma_k \tau_k^{\alpha_i} v_i(t) + \sum_{k=1}^{l} \gamma_k \tau_k^{\alpha_l} \eta^{\tau_k}(t).$$

Etant données les égalités (4.16) et l'indépendance des fonctions u et v_i par rapport à τ_k,

$$U^H(t) = u(t) + \sum_{k=1}^{l} \gamma_k \tau_k^{\alpha_l} \eta^{\tau_k}(t).$$

D'où

$$|U^H(t) - u(t)| \leqslant \sum_{k=1}^{l} |\gamma_k| \tau_k^{\alpha_l} |\eta^{\tau_k}(t)|.$$

On évalue le second membre à l'aide de (4.13) et de la définition de τ_k:

$$|U^H(t) - u(t)| \leqslant \tau^{\alpha_l} c_1 \sum_{k=1}^{l} \frac{|\gamma_k|}{N_k^{\alpha_l}}.$$

Comme c_1, γ_k et N_k ne dépendent pas de τ et M, cette inégalité montre que la précision obtenue sur U^H est de l'ordre de τ^{α_l} pour $M \to \infty$.

On se place, pour concrétiser, dans le problème (4.4), (4.5) et on vérifie l'efficacité de l'extrapolation basée sur le développement suivant les puissances fractionnaires. On suppose que la solution de (4.4), (4.5) admet le développement (4.12), où $l = 2$, $\alpha_1 = 3/2$,

$\alpha_2 = 2$. On pose $N_1 = 1$, $N_2 = 2$ et on cherche pour plusieurs M entiers naturels les solutions u^{τ_k} associées aux réseaux de pas $\tau_k = 1/(N_k M)$. L'erreur $\psi^\tau(1)$ sur u^{τ_k} est donnée par la fig. 2.2. Le système (4.15) s'écrit dans notre cas

$$\gamma_1 + \gamma_2 = 1,$$
$$\gamma_1 + 2^{-3/2}\gamma_2 = 0.$$

et possède comme solution

$$\gamma_1 = -\frac{2^{-3/2}}{1 - 2^{-3/2}}, \qquad \gamma_2 = \frac{1}{1 - 2^{-3/2}}.$$

On forme la solution extrapolée U^H pour plusieurs M entiers naturels:

$$U^H(t) = \gamma_1 u^\tau(t) + \gamma_2 u^{\tau/2}(t), \quad \text{où} \quad \tau = 1/M.$$

On calcule l'erreur

$$\eta_3(3M) = \left| U^H(1) - u(1) \right|. \tag{4.17}$$

On l'a rapportée à 3 M car le calcul de U^H exige sensiblement le même nombre d'opérations que celui de la solution $u^{\tau/3}$.

La pente du graphe de η_3 confirme l'ordre de précision élevé des solutions approchées des problèmes avec singularités qu'on obtient par extrapolation (fig. 2.2).

L'intégration d'une équation différentielle ordinaire du premier ordre est étroitement liée avec les formules de quadrature, si bien qu'en théorie des quadratures on raffine souvent les résultats à l'aide des développements suivant les puissances fractionnaires de τ (et des fois suivant celles de $\tau^\alpha \ln^p \tau$). Pour ces questions voir [22], [66], [86], [87], [94], [131], où l'on trouve de plus des résultats de la résolution numérique d'équations différentielles ordinaires avec singularités.

La méthode que nous venons de décrire paraît avoir un vaste avenir en ce qui concerne les équations différentielles ordinaires et les équations aux dérivées partielles (pour celles-ci surtout). En effet, les résultats pratiques sont prometteurs, même pour le problème de valeurs propres en dimension deux (voir [85]). On n'arrive par contre que fort rarement à justifier rigoureusement les développements de la forme (4.12). Un autre inconvénient de la méthode est qu'elle est difficilement mécanisable car elle ne débute qu'après des calculs analytiques laborieux pour établir l'ordre des α_i.

On considérera plus loin (ch. 4) une autre méthode pour les équations aux dérivées partielles. Elle repose sur le morcellement du réseau au voisinage de la singularité et permet de résoudre numériquement des problèmes dont les solutions présentent des singularités ponctuelles.

CHAPITRE 3

ÉQUATION DE DIFFUSION STATIONNAIRE EN DIMENSION UN

L'intérêt de l'équation de diffusion tient à ce qu'elle trouve de nombreuses applications dans diverses branches scientifiques. Elle occupe une place de premier ordre en théorie des réacteurs nucléaires où l'approximation de diffusion de l'équation de transport a joué un rôle hors pair. La protection de l'environnement relève du calcul de la diffusion des aérosols industriels. Des problèmes de diffusion d'une grande portée se posent en physique, en chimie, en géophysique... Aussi ce chapitre sera centré sur plusieurs façons de poser les problèmes caractéristiques des équations de diffusion.

Nous tenons à noter que s'agissant des applications particulièrement intéressantes, l'équation de diffusion intervient lorsque les coefficients et la fonction de sources sont discontinus. Cela impose des restrictions supplémentaires sur la solution qui se trouve continue sans être dérivable. On doit donc modifier sensiblement l'appareil mathématique développé dans le Chapitre 2, qui permet de trouver les solutions améliorées dans leur ensemble à partir des solutions approchées associées à des réseaux différents. En effet, le développement taylorien de la solution a lieu seulement dans les domaines privés de points où les coefficients et la fonction de sources subissent des discontinuités. N'empêche que les problèmes à coefficients continûment dérivables présentent un intérêt certain, si bien que notre étude de l'extrapolation de Richardson pour les problèmes aux limites débutera par l'équation de diffusion stationnaire à coefficients suffisamment réguliers.

3.1. Problème de Dirichlet

Dans ce paragraphe, on étudiera le problème de Dirichlet pour l'équation de diffusion. Contrairement aux chapitres précédents, où l'on s'est occupé seulement du problème de Cauchy pour des équations différentielles ordinaires, on insistera cette fois sur les problèmes aux limites pour des équations différentielles du second ordre dont les équations de diffusion constituent le cas le plus intéressant du point de vue des applications. On se place d'abord dans le cas de coefficients assez réguliers et on justifie une méthode

d'extrapolation. Comme plus haut, on obtient une solution approchée précise au maximum à l'aide des solutions approchées des équations aux différences d'ordre d'approximation peu élevé, qui sont associées à une suite de réseaux.

Ainsi, soit le problème de Dirichlet en dimension un pour l'équation de diffusion stationnaire

$$-(pu')' + qu = f, \qquad x \in (0, 1), \tag{1.1}$$

$$u(0) = u_0, \qquad u(1) = u_1. \tag{1.2}$$

En ce qui concerne les coefficients du problème, on suppose que

$$p(x) \geqslant c_1 > 0, \qquad q(x) \geqslant 0, \qquad x \in (0, 1), \tag{1.3}$$

et

$$q, f \in C^r[0, 1], \qquad p \in C^{r+1}[0, 1] \tag{1.4}$$

pour un entier naturel $r \geqslant 2$.

Théorème 1.1. *On suppose que le problème* (1.1), (1.2) *vérifie les conditions* (1.3), (1.4). *Il existe une solution unique* $u \in C^{r+2}[0, 1]$.

Démonstration. On cherche la solution sous forme de somme

$$u(x) = v(x) + \big(u_0 + x(u_1 - u_0)\big),$$

v étant solution du problème

$$\begin{aligned} -(pv')' + qv = g \quad \text{sur} \quad (0, 1) \\ v(0) = v(1) = 0, \end{aligned} \tag{1.5}$$

où $g = f + (u_1 - u_0)p - q\big(u_0 + x(u_1 - u_0)\big)$. Il existe (voir [92]) une fonction de Green $G(x, t)$ telle que

$$v(x) = \int_0^1 G(x, t)\, g(t)\, dt.$$

La continuité de G entraîne la même propriété sur $[0, 1]$ de $v(x)$. On porte le terme continu qv dans le second membre de (1.5) et on intègre de t à x, il vient

$$-p(x)v'(x) + p(t)v'(t) = -\int_t^x \big(q(\xi)v(\xi) - g(\xi)\big)\, d\xi. \tag{1.6}$$

On divise par $-p(t)$ et on intègre par rapport à t de 0 à x:

$$p(x)v'(x)\int_0^x \frac{dt}{p(t)} - v(x) = \int_0^x \frac{dt}{p(t)} \int_t^x \big(q(\xi)v(\xi) - g(\xi)\big)\, d\xi.$$

On en tire v' :

$$v'(x) = \frac{1}{p(x)} \frac{1}{\int\limits_0^x \frac{dt}{p(t)}} \left\{ v(x) + \int\limits_0^x \frac{1}{p(t)} \int\limits_t^x \left(q(\xi)\, v(\xi) - g(\xi) \right) d\xi \quad dt \right\}.$$

Plaçons-nous dans le cas du segment [1/3, 1]. Toutes les fonctions sous $\int$ sont continues, et $p(t)$ est strictement positive, si bien que l'expression dans l'accolade est continue. La fonction

$$\int\limits_0^x \frac{dt}{p(t)}$$

est elle aussi continue en x et strictement positive sur [1/3, 1]. Aussi l'inverse jouit également de la propriété de continuité. Ainsi, la dérivée v' est continue sur [1/3, 1] en tant que produit de fonctions continues.

Reprenons l'identité (1.6). On divise par $p(t)$ et on intègre de x à 1. L'égalité ainsi obtenue entraîne la relation

$$v'(x) = \frac{1}{p(x)} \frac{1}{\int\limits_x^1 \frac{dt}{p(t)}} \left\{ -v(x) + \int\limits_x^1 \frac{1}{p(t)} \int\limits_t^x \left(q(\xi)\, v(\xi) - g(\xi) \right) d\xi \quad dt \right\}.$$

D'où la continuité de v' sur [0, 2/3], si bien qu'elle l'est sur [0, 1] tout entier et $v \in C^1$ [0, 1]. On a par des transformations analogues de (1.5):

$$v'' = -\frac{p'}{p} v' + \frac{q}{p} v - \frac{g}{p}. \tag{1.7}$$

Tous les termes du second membre étant continus, il en est de même de v'' sur [0, 1]. Aussi $v \in C^2$ [0, 1], ce qui entraîne à son tour la propriété des termes du second membre de (1.7) d'être continûment dérivables. C'est pourquoi, v'' l'est sur [0, 1], donc $v \in$ $\in C^3$ [0, 1]. Au bout de $r - 1$ autres pas, on obtient le résultat $v \in C^{r+2}[0, 1]$.

Comme u est la somme de la fonction v et d'un polynôme, on a le résultat de régularité voulu. L'unicité découle de façon classique du principe du maximum (voir par exemple [58]). Le théorème se trouve démontré.

On résout numériquement le problème proposé en construisant le réseau régulier

$$\bar{\omega}_h = \{ x_i = ih, \quad i = 0, 1, \ldots, N \}$$

de pas $h = 1/N$, N étant un entier naturel. On désigne par $\bar{\omega}_h$ l'ensemble des points médians

$$\bar{\omega}_h = \{ x_{i+1/2} = (i + 1/2)\, h, \quad i = 0, 1, \ldots, N-1 \}$$

et par ω_h l'ensemble des points intérieurs

$$\omega_h = \{ x_i = ih, \quad i = 1, \ldots, N-1 \}.$$

On fait correspondre à chaque point une équation aux différences

$$-(pu^h_{\bar{x}})_{\hat{x}} + qu^h = f \quad \text{sur} \quad \omega_h. \tag{1.8}$$

On a, par définition, pour v arbitraire

$$v_{\bar{x}} = \frac{v(x + h/2) - v(x - h/2)}{h},$$

si bien que

$$(pv_{\bar{x}})_{\hat{x}\,|\,x} = \frac{p(x + h/2)\big(v(x+h) - v(x)\big) - p(x - h/2)\big(v(x) - v(x-h)\big)}{h^2}.$$

L'équation (1.8) plus deux conditions aux limites

$$u^h(0) = u_0, \qquad u^h(1) = u_1 \tag{1.9}$$

donnent $N+1$ équations pour déterminer $N+1$ valeurs de la fonction discrète u^h aux nœuds du réseau $\bar{\omega}_h$.

Le lemme 1.2, § 7.1 entraîne que, sous les hypothèses (1.4), le problème aux différences est approché d'ordre 2. Voyons ce qu'il en est de la stabilité.

LEMME 1.2. *On suppose que le problème aux différences* (1.8), (1.9) *vérifie les conditions* (1.3). *La solution du problème admet l'estimation*

$$\| u^h \|_{C,h} \leqslant \frac{1}{c_1} \max_{\omega_h} |f| + \max \{ |u(0)|, |u(1)| \}.$$

DÉMONSTRATION. Mettons la solution sous forme de somme de deux termes: $u^h = v^h + w^h$ qui sont solutions respectives des problèmes

$$-(pv^h_{\bar{x}})_{\hat{x}} + qv^h = 0 \quad \text{sur} \quad \omega_h,$$
$$v^h(0) = u(0), \qquad v^h(1) = u(1)$$

et

$$-(pw^h_{\bar{x}})_{\hat{x}} + qw^h = f \quad \text{sur} \quad \omega_h,$$
$$w^h(0) = w^h(1) = 0$$

Le premier problème vérifie le principe du maximum sous forme discrète (voir [43]) qui implique la majoration

$$\| v^h \|_{C,h} \leq \max \{| u(0) |; | u(1) |\}.$$

Quant au second, on emprunte à [43]

$$\| w^h \|_{C,h} \leq \frac{1}{c_1} \max_{\omega_h} | f |.$$

Les deux estimations et l'inégalité du triangle

$$\| u^h \|_{C,h} \leq \| v^h \|_{C,h} + \| w^h \|_{C,h}$$

garantissent la majoration voulue.

On utilise l'affirmation du lemme et le fait que (1.8), (1.9) est un schéma d'approximation d'ordre 2, et on démontre que u^h est exacte à l'ordre 2 en h.

On construit une solution corrigée plus précise. A cet effet, on pose $l = [(r-1)/2]$ et on fixe $M_1 < \ldots < M_{l+1}$ entiers naturels. Le problème (1.8), (1.9) est résolu sur chaque réseau $\bar{\omega}_{h_k}$, $h_k = 1/(M_k N)$.

N est une fois de plus supposé indéfiniment croissant, et on demande la relation entre la précision de la solution corrigée et le paramètre $h = 1/N$. On note que toutes les solutions u^{h_k} sont définies sur le réseau $\bar{\omega}_h$.

Il est déjà apparu que le système

$$\begin{aligned} &\sum_{k=1}^{l+1} \gamma_k = 1, \\ &\sum_{k=1}^{l+1} \gamma_k h_k^{2j} = 0, \qquad j = 1, \ldots, l, \end{aligned} \tag{1.10}$$

est non dégénéré. Prenons sa solution unique $\gamma_1, \ldots, \gamma_{l+1}$ et formons la combinaison linéaire

$$U^H(x) = \sum_{k=1}^{l+1} \gamma_k u^{h_k}(x), \qquad x \in \bar{\omega}_h. \tag{1.11}$$

THÉORÈME 1.3. *On suppose que le problème* (1.1), (1.2) *satisfait aux conditions* (1.3), (1.4). *La solution corrigée* (1.11) *avec les poids fournis par le système* (1.10) *admet l'estimation*

$$\| U^H - u \|_{C,h} \leq c_2 h^r. \tag{1.12}$$

Démonstration. Voyons si l'on est dans les hypothèses du théorème 2.2, § 1.2. On pose $M_k(\Omega) = C^k\,[0, 1]$, $P_k(\bar\Omega) = C^{k+2}[0, 1]$ et $N_k(D) = \mathbf{R}^2$. La condition A du § 1.2 est une conséquence du théorème 1.1. On pose pour le problème aux différences (2.3) du paragraphe mentionné:

$$\bar\Omega_h = \bar\omega_h, \qquad \breve\Omega_h = \omega_h, \qquad D_h = \{0, 1\}$$

et

$$\| u \|_{\bar\Omega_h} = \| u \|_{C,h}, \qquad \| u \|_{\breve\Omega_h} = \max_{\omega_h} | u |,$$

$$\| u \|_{D_h} = \max \{| u(0) |, | u(1) |\}.$$

Dans ce cas, la condition B du § 1.2 coïncide avec l'estimation à priori du lemme 1.2. Voyons si l'erreur d'approximation est développable suivant les puissances paires de h. Le lemme 1.2, § 7.1 implique pour toute fonction $\varphi \in C^{r-2k+2}\,[0, 1]$ le développement

$$-(p\,\varphi_{\bar x})_x + q\,\varphi = -(p\,\varphi')' + q\,\varphi -$$

$$-\sum_{j=1}^{l-k} h^{2j}\,4^{-j} \sum_{k+s=j} \frac{(p\,\varphi^{(2k+1)})^{(2s+1)}}{(2k+1)!\,(2s+1)!} + h^{r-2k}\,\sigma^h \quad \text{sur} \quad \omega_h,$$

et $|\sigma^h| \leqslant c_3 \quad \forall x \in \omega_h$.

Ainsi, la condition D est remplie pour la constante $\beta = 0$. On a toutes les conditions du théorème 2.2, § 1.2. On passe au théorème 3.2, § 1.3. La condition (3.15) de son énoncé est vérifiée avec la constante

$$d_3 = \min_{1 \leqslant k \leqslant l} \left(\frac{M_{k+1}}{M_k}\right) - 1.$$

Le théorème entraîne donc l'estimation

$$\max_{\bar\Omega_h} | U^H - u | \leqslant d_4\, h_1^r,$$

où d_4 est une constante indépendante de h_k. Comme l'intersection $\bar\Omega_H$ contient $\bar\omega_h$, on a

$$\| U^H - u \|_{C,h} \leqslant \frac{d_4}{M_1^r}\, h^r.$$

On pose $c_2 = d_4/M_1^r$, il vient (1.12), ce qui démontre le théorème 1.3.

Ainsi, la précision de la solution corrigée dépend seulement de l'indice de régularité r des données du problème (1.1), (1.2).

Illustrons les résultats théoriques par un exemple numérique. Soit le problème

$$-((1+x)u')' + xu = \frac{1+x^2+x^3}{(1+x)^2} \text{ sur } (0,\ 1),$$
$$u(0) = 0, \qquad u(1) = 1/2.$$

Sa solution analytique s'écrit

$$u(x) = \frac{x}{1+x}.$$

Ci-dessous le tableau des erreurs maxima en fonction du nombre de nœuds du réseau $\bar{\omega}_h$.

Tableau 3.1

N	Erreur maximum sur la solution du problème (1.8), (1.9)	Erreur maximum sur la solution extrapolée (1.11)		
		$h_1 = 1/N$, $h_2 = 1/(2N)$	$h_1 = 1/N$, $h_2 = 1/(2N)$, $h_3 = 1/(3N)$	$h_1 = 1/N$, $h_2 = 1/(2N)$, $h_3 = 1/(3N)$, $h_4 = 1/(4N)$
10	$2{,}3\cdot10^{-4}$	$4{,}2\cdot10^{-7}$	$5{,}3\cdot10^{-10}$	$5{,}6\cdot10^{-13}$
20	$5{,}8\cdot10^{-5}$	$2{,}6\cdot10^{-8}$	$8{,}4\cdot10^{-12}$	$4{,}6\cdot10^{-15}$
30	$2{,}6\cdot10^{-5}$	$5{,}3\cdot10^{-9}$	$7{,}5\cdot10^{-13}$	$2{,}1\cdot10^{-14}$
40	$1{,}4\cdot10^{-5}$	$1{,}7\cdot10^{-9}$	$1{,}5\cdot10^{-13}$	$6.3\cdot10^{-13}$

On note que les erreurs relatives d'arrondi se situent au niveau de 10^{-14}, ce qui se voit bien dans deux dernières colonnes.

3.2. Troisième problème aux limites

On se heurte dans le troisième problème aux limites à une difficulté absente du cas de Dirichlet relatif à l'équation de diffusion. Il s'agit du problème d'approcher les conditions aux limites. Ce paragraphe se propose de justifier la méthode de Richardson pour le troisième problème aux limites.

Soit, pour l'équation (1.1) sous les hypothèses (1.3), (1.4), le troisième problème aux limites

$$\alpha_0 u(0) - u'(0) = g_0, \tag{2.1}$$

$$\alpha_1 u(1) + u'(1) = g_1, \tag{2.2}$$

avec α_0 et α_1 des constantes non négatives telles que

$$\alpha_0 + \alpha_1 = c_2 > 0. \tag{2.3}$$

Théorème 2.1. *On suppose que le problème* (1.1), (2.1), (2.2) *vérifie les conditions* (1.3), (1.4), (2.3). *Il existe une solution unique* $u \in C^{r+2}\,[0,\ 1]$.

Démonstration. On met la solution u sous forme de somme: $u(x) = v(x) + ax + b$, les constantes a et b étant définies par les égalités

$$b = \frac{(1+\alpha_1)\,g_0 + g_1}{\alpha_0 + \alpha_1 + \alpha_0\,\alpha_1}, \qquad a = \alpha_0\, b - g_0.$$

Ce choix de a et b garantit le caractère homogène de la condition aux limites pour v:

$$\alpha_0\, v - v' = 0 \qquad \text{au point } x = 0,$$
$$\alpha_1\, v + v' = 0 \qquad \text{au point } x = 1.$$

La fonction v est de plus telle que

$$-(p\,v')' + qv = g \qquad \text{sur } [0,\ 1],$$

où

$$g = f + p'\,a - q\,(ax + b).$$

La suite coïncide textuellement avec la démonstration du théorème 1.1 jusqu'au résultat de [92] qui reste valable pour le troisième problème aux limites.

Si l'on construit le schéma aux différences à l'aide de l'équation (1.8) associée aux nœuds de ω_h (comme c'est par exemple le cas de [43]), on approche les conditions aux limites par des différences unilatérales. On conçoit que cela équivaut à l'impossibilité de développer suivant les puissances paires de h. En effet, les puissances impaires de h ne disparaissent automatiquement que si l'on utilise les différences centrales, et les développements à coefficients non nuls des puissances impaires sont moins efficaces parce qu'on a à éliminer dans l'extrapolation deux fois plus de termes, i.e. on a à résoudre deux fois plus de problèmes approchés.

Aussi on propose dans [24] un schéma basé sur les équations discrètes

$$-(p u^h_{\bar x})_x + q u^h = f \quad \text{sur} \quad \breve{\omega}_h, \tag{2.4}$$

et on fait disparaître dans la première et la dernière équation les points extérieurs à [0, 1]. On utilise à cet effet les équations approchant les conditions aux limites (2.1), (2.2)

$$\alpha_0\, u^h_{\breve{x}} - u^h_{\mathring{x}} = g_0 \qquad \text{au point } x = 0, \tag{2.5}$$

$$\alpha_1\, u^h_{\breve{x}} + u^h_{\mathring{x}} = g_1 \qquad \text{au point } x = 1. \tag{2.6}$$

On aboutit par élimination à un système algébrique de matrice $N \times N$ symétrique. Si l'on se passe de l'élimination, la matrice

du système n'est pas symétrique, ce qui n'a pas (dans le cas de matrices tridiagonales) d'influence décisive sur le procédé de résolution (la méthode du balayage). D'autre part, la méthode d'extrapolation devient beaucoup plus intuitive parce qu'avec le procédé sans élimination des inconnues, on apprécie séparément l'apport des approximations des conditions aux limites et celui des approximations de l'équation même.

Ainsi, on se propose de résoudre numériquement le système (2.4) à (2.6). On prolonge au préalable la fonction $u(x)$ en dehors de $[0, 1]$, disons, par les polynômes

$$\sum_{k=0}^{r+2} \frac{x^k}{k!} u^{(k)}(0), \qquad \sum_{k=0}^{r+2} \frac{(x-1)^k}{k!} u^{(k)}(1)$$

à gauche et à droite respectivement. La solution prolongée est de classe C^{r+2} $[-1/2, 3/2]$, et on la désignera par la notation usuelle $u(x)$.

On vérifie que la solution du schéma jouit de la propriété de stabilité.

THÉORÈME 2.2. *La solution du système* (2.4) *à* (2.6) *vérifie l'estimation*

$$\max_{x \in \bar{\omega}_h} |u^h(x)| \leqslant \frac{2(2+c_2)}{c_1 c_2} \left(\max_{x \in \bar{\omega}_h} |f(x)| + p(0)|g_0| + p(1)|g_1|\right).$$

DÉMONSTRATION. On fait le produit de chaque équation (2.4) associée au point $x_{i+1/2} \in \bar{\omega}_h$ par $hu^h(x_{i+1/2})$ et on additionne les résultats, il vient

$$-\sum_{x \in \bar{\omega}_h} (pu^h_{\bar{x}})_x u^h h + \sum_{x \in \bar{\omega}_h} q(u^h)^2 h = \sum_{x \in \bar{\omega}_h} fu^h h. \tag{2.7}$$

On applique au premier terme la première formule de Green discrète (voir [41]) qui s'écrit avec nos notations

$$-\sum_{x \in \bar{\omega}_h} v_{\bar{x}}(x) w(x) h =$$
$$= \sum_{x \in \omega_h} v(x) w_{\bar{x}}(x) h + v(0) w(h/2) - v(1) w(1-h/2),$$

la fonction v étant définie aux nœuds de ω_h et la fonction $w(x)$ en ceux de $\bar{\omega}_h$. On trouve l'égalité

$$-\sum_{\bar{\omega}_h} (pu^h_{\bar{x}})_x u^h h = \sum_{x \in \omega_h} p(x)(u^h_{\bar{x}}(x))^2 h -$$
$$- p(1) u^h(1-h/2) u^h_{\bar{x}}(1) + p(0) u^h(h/2) u^h_{\bar{x}}(0). \tag{2.8}$$

On récrit (2.5) et (2.6):

$$-\left(1+\frac{h\alpha_0}{2}\right)u_{\bar x}^h(0)+\alpha_0 u^h(h/2)=g_0,$$

$$\left(1+\frac{h\alpha_1}{2}\right)u_{\bar x}^h(1)+\alpha_1 u^h(1-h/2)=g_1,$$

et on met (2.8) sous forme

$$-\sum_{\bar\omega_h}(p\,u_{\bar x}^h)_{\hat x}\,u^h\,h=$$

$$=\sum_{x\in\omega_h}p(x)\,(u_{\bar x}^h(x))^2 h+\frac{p(0)}{1+h\alpha_0/2}\left(\alpha_0 u^h\left(\frac{h}{2}\right)-g_0\right)u^h\left(\frac{h}{2}\right)+$$

$$+\frac{p(1)}{1+h\alpha_1/2}\left(\alpha_1 u^h\left(1-\frac{h}{2}\right)-g_1\right)u^h\left(1-\frac{h}{2}\right).$$

Par substitution dans la formule (2.7), on a après certaines transformations

$$\sum_{x\in\omega_h}p(x)\,(u_{\bar x}^h(x))^2 h+\frac{\alpha_0 p(0)}{1+h\alpha_0/2}\left(u^h\left(\frac{h}{2}\right)\right)^2+$$

$$+\frac{\alpha_1 p(1)}{1+h\alpha_1/2}\left(u^h\left(1-\frac{h}{2}\right)\right)^2+\sum_{x\in\bar\omega_h}q(x)\,(u^h(x))^2 h=$$

$$=\sum_{x\in\bar\omega_h}f(x)\,u^h(x)\,h+\frac{p(0)\,g_0}{1+h\alpha_0/2}u^h\left(\frac{h}{2}\right)+\frac{p(1)\,g_1}{1+h\alpha_1/2}u^h\left(1-\frac{h}{2}\right). \quad (2.9)$$

On minore ou on majore chaque terme de l'égalité (2.9) selon qu'il fait partie du premier ou du second membre. Etant donné $p(x)\geqslant c_1$ (propriété (1.3) du paragraphe précédent), on a l'inégalité

$$\sum_{x\in\omega_h}p(x)\,(u_{\bar x}^h(x))^2 h\geqslant c_1\sum_{x\in\omega_h}(u_{\bar x}^h(x))^2 h. \quad (2.10)$$

Comme $h<1$, $\alpha_i\geqslant 0$, on a

$$\frac{\alpha_0 p(0)}{1+h\alpha_0/2}\left(u^h\left(\frac{h}{2}\right)\right)^2\geqslant\frac{\alpha_0 c_1}{1+\alpha_0/2}\left(u^h\left(\frac{h}{2}\right)\right)^2. \quad (2.11)$$

$$\frac{\alpha_1 p(1)}{1+h\alpha_1/2}\left(u^h\left(1-\frac{h}{2}\right)\right)^2\geqslant\frac{\alpha_1 c_1}{1+\alpha_1/2}\left(u^h\left(1-\frac{h}{2}\right)\right)^2. \quad (2.12)$$

D'après la propriété (1.3) mentionnée, $q(x)\geqslant 0$, si bien que

$$\sum_{x\in\bar\omega_h}q(x)\,(u^h(x))^2 h\geqslant 0. \quad (2.13)$$

L'estimation

$$\sum_{x\in\bar\omega_h} f(x)\, u^h(x)\, h \leqslant \max_{x\in\bar\omega_h} |f(x)| \max_{x\in\bar\omega_h} |u^h(x)| \tag{2.14}$$

est obtenue en remplaçant chaque terme par son maximum et en prenant en considération que les points de $\bar\omega_h$ sont au nombre de N et que $hN = 1$. Les inégalités simples

$$\frac{p(0)\, g_0}{1+\alpha_0 h/2}\, u^h\left(\frac{h}{2}\right) \leqslant p(0)\, |g_0| \max_{x\in\bar\omega_h} |u^h(x)|. \tag{2.15}$$

$$\frac{p(1)\, g_1}{1+h\,\alpha_1/2}\, u^h\left(1-\frac{h}{2}\right) \leqslant p(1)\, |g_1| \max_{x\in\bar\omega_h} |u^h(x)| \tag{2.16}$$

découlent de $\alpha_i \geqslant 0$, $h > 0$.

La relation (2.9) et les inégalités (2.10) à (2.16) donnent l'estimation

$$c_1 \sum_{x\in\omega_h} \left((u_{\bar x}^h(x))^2 h + \frac{\alpha_0 c_1}{1+\alpha_0/2} \left(u^h\left(\frac{h}{2}\right)\right)^2 + \frac{\alpha_1 c_1}{1+\alpha_1/2} \left(u^h\left(1-\frac{h}{2}\right)\right)^2 \leqslant \right.$$

$$\leqslant (\max_{x\in\bar\omega_h} |f(x)| + p(0)\,|g_0| + p(1)\,|g_1|) \max_{x\in\bar\omega_h} |u^h(x)|. \tag{2.17}$$

On se rappelle que $\alpha_0 + \alpha_1 \geqslant c_2$. Il est clair que l'un de ces nombres est au moins égal à $c_2/2$. On estime, sans restreindre la généralité, que $c_2 \geqslant \alpha_0 \geqslant c_2/2$. Aussi

$$\frac{\alpha_0}{1+\alpha_0/2} \geqslant \frac{c_2}{2+c_2}.$$

On a de plus

$$\frac{c_2}{2+c_2} < 1, \qquad \frac{\alpha_1}{1+\alpha_1/2} \geqslant 0.$$

On s'en sert pour évaluer le premier membre de (2.17), il vient

$$\frac{c_1 c_2}{2+c_2}\left(\sum_{x\in\omega_h} (u_{\bar x}^h(x))^2 h + \left(u^h\left(\frac{h}{2}\right)\right)^2\right) \leqslant$$

$$\leqslant (\max_{x\in\bar\omega_h} |f(x)| + p(0)|\, g_0\,| + p(1)|\, g_1\,|) \max_{x\in\bar\omega_h} |\, u^h(x)|.$$

On simplifie l'inégalité en utilisant l'estimation discrète de l'immersion (voir [41]) (un analogue de l'estimation des normes dans $C[0, 1]$ et $W_2^1(0, 1)$)

$$\max_{x\in\bar\omega_h} (u^h(x))^2 \leqslant 2\left(\sum_{x\in\omega_h} (u_{\bar x}^h(x))^2 h + (u^h(h/2))^2\right)$$

qui donne

$$\frac{c_1 c_2}{2(2+c_2)} \max_{\bar{\omega}_h} (u^h)^2 \leqslant \frac{c_1 c_2}{2+c_2}\left(\sum_{\omega_h} (u^h_{\bar{x}})^2 h + \left(u^h\left(\frac{h}{2}\right)\right)^2\right).$$

ce qui permet de récrire (2.17)

$$\frac{c_1 c_2}{2(2+c_2)} \max_{\bar{\omega}_h} | u^h |^2 \leqslant \max_{\bar{\omega}_h} | u^h | (\max_{\bar{\omega}_h} | f | + p(0) | g_0 | + p(1) | g_1 |).$$

L'affirmation du théorème 2.2 résulte de cette inégalité divisée membre à membre par

$$\frac{c_1 c_2}{2(2+c_2)} \max_{\bar{\omega}_h} | u^h |.$$

Avec l'estimation ainsi démontrée, on dit que le problème (2.4) à (2.6) admet une solution unique pour chaque réseau $\bar{\omega}_h$.

La construction de la solution corrigée à partir de plusieurs solutions connues obéit à la règle suivante. On pose $l = [(r-1)/2]$ et on fixe les nombres impairs $0 < M_1 \ldots < M_{l+1}$. On résout pour chaque réseau $\bar{\omega}_{h_k}$, $h_k = 1/(M_k N)$, le problème aux différences (2.4) à (2.6). On note qu'avec M_k impairs, on a pour tout entier naturel N l'inclusion $\bar{\omega}_{h_k} \supset \bar{\omega}_h$ $(h = 1/N)$; aussi toutes les solutions sont définies sur $\bar{\omega}_h$.

Soit le système

$$\begin{aligned} \sum_{k=1}^{l+1} \gamma_k &= 1, \\ \sum_{k=1}^{l+1} \gamma_k h_k^{2j} &= 0, \qquad j = 1, \ldots, l. \end{aligned} \tag{2.18}$$

Il possède pour seule solution $\gamma_1, \ldots, \gamma_{l+1}$. On forme la combinaison linéaire

$$U^H(x) = \sum_{k=1}^{l+1} \gamma_k u^{h_k}(x), \qquad x \in \bar{\omega}_h, \tag{2.19}$$

et on établit son ordre de précision.

THÉORÈME 2.3. *On suppose que le problème différentiel* (1.1), (2.1), (2.2) *remplit les conditions* (1.3), (1.4), (2.3). *La solution corrigée* (2.19) *avec les poids définis à partir du système* (2.18) *admet l'estimation*

$$\max_{\bar{\omega}_h} | U^H - u | \leqslant c_3 h^r. \tag{2.20}$$

DÉMONSTRATION. Vérifions si les hypothèses du théorème 2.2, § 1.2, sont bien satisfaites. On pose $M_k(\Omega) = C^k[0, 1]$, $P_k(\bar{\Omega}) = C^{k+2}[0, 1]$ et $N_k(D) = \mathbf{R}^2$ pour tout k. La condition A du § 1.2 découle dans ce cas du théorème 2.1. On pose pour le problème aux différences (2.3) du § 1.2: $\bar{\Omega}_h = \bar{\omega}_h$, $\mathring{\Omega}_h = \bar{\omega}_h$, $D_h = \{0, 1\}$ et

$$\|u\|_{\bar{\Omega}_h} = \|u\|_{\mathring{\Omega}_h} = \max_{\bar{\omega}_h} |u|,$$

$$\|u\|_{D_h} = \max\{|u(0)|, |u(1)|\}.$$

Avec ces notations, la condition B du § 1.2 coïncide avec l'estimation du théorème 2.2. Il ne nous reste qu'à établir si les erreurs d'approximation se développent suivant les puissances paires de h. Par le lemme 1.2, § 7.1, toute fonction $\varphi \in C^{r-2k+2}[0, 1]$ admet le développement

$$-(p\varphi_{\bar{x}})_x + q\varphi = -(p\varphi')' + q\varphi -$$
$$-\sum_{j=1}^{l-k} h^{2j} 4^{-j} \sum_{k+s=j} \frac{(p\varphi^{(2k+1)})^{(2s+1)}}{(2k+1)!\,(2s+1)!} + h^{r-2k}\sigma^h \text{ sur } \omega_h,$$

et

$$|\sigma^h| \leqslant c_4 \qquad \forall x \in \omega_h.$$

On a aux points frontières

$$\alpha_0 \varphi_{\bar{x}}(0) - \varphi_x(0) = \alpha_0 \varphi(0) - \varphi'(0) +$$
$$+\sum_{j=1}^{l-k} h^{2j}4^{-j}\left(\alpha_0 \frac{\varphi^{(2j)}(0)}{(2j)!} - \frac{\varphi^{(2j+1)}(0)}{(2j+1)!}\right) + h^{r-2k}\rho^0,$$

où $|\rho_0| \leqslant c_5$, et

$$\alpha_1 \varphi_{\bar{x}}(1) + \varphi_{\bar{x}}(1) = \alpha_1 \varphi(1) + \varphi'(1) +$$
$$+\sum_{j=1}^{l-k} h^{2j}4^{-j}\left(\alpha_1 \frac{\varphi^{(2j)}(1)}{(2j)!} + \frac{\varphi^{(2j+1)}(1)}{(2j+1)!}\right) + h^{r-2k}\rho_1,$$

où $|\rho_1| \leqslant c_6$.

Ainsi, la condition D est vérifiée avec la constante $\beta = 0$. On est donc dans toutes les hypothèses du théorème 2.2, § 1.2. Voyons ce qu'il en est de celles du théorème 3.2, § 1.3. Il est immédiat d'établir qu'on n'a à tester que (3.15). Cette condition devient évidente si l'on prend

$$d_3 = \min_{1 \leqslant k < l} \left(\frac{M_{k+1}}{M_k}\right) - 1.$$

Aussi le théorème 3.2 entraîne l'estimation

$$\max_{\bar{\Omega}_h} | U^H - u | \leq d_4 h_1^r,$$

la constante d_4 étant indépendante de h_k. Les nombres M_k sont impairs, si bien que l'intersection $\bar{\Omega}_H$ contient $\bar{\omega}_h$, donc

$$\max_{\bar{\omega}_h} | U^H - u | \leq \frac{d_4}{M_1^r} h^r.$$

On aboutit au résultat désiré (2.20) si l'on prend $c_3 = d_4/M_1^r$.

Ainsi, on a établi pour le troisième problème aux limites que la précision du correcteur linéaire dépend seulement de l'indice de régularité r des coefficients et du second membre de l'équation (1.1).

Pour illustrer le gain en précision apporté par le procédé décrit, on considère le problème

$$-((1 + x) u')' + xu = \frac{1 + x^2 + x^3}{(1 + x)^2} \text{ sur } (0, 1),$$

$$u(0) - u'(0) = -1, \qquad 2u(1) + u'(1) = 5/4.$$

Sa solution analytique s'écrit

$$u(x) = \frac{x}{1 + x}.$$

Le tableau 3.2 donne les erreurs maxima en fonction du nombre de nœuds de $\bar{\omega}_h$.

Tableau 3.2

N	Erreur maximum sur la solution du problème (2.4) à (2.6)	Erreur maximum sur la solution extrapolée (2.19)		
		$h_1 = 1/N$, $h_2 = 1/(3N)$	$h_1 = 1/N$, $h_2 = 1/(3N)$, $h_3 = 1/(5N)$	$h_1 = 1/N$, $h_2 = 1/(3N)$, $h_3 = 1/(5N)$, $h_4 = 1/(7N)$
10	$3,5 \cdot 10^{-3}$	$7,7 \cdot 10^{-7}$	$8,6 \cdot 10^{-10}$	$7,6 \cdot 10^{-13}$
20	$8,7 \cdot 10^{-4}$	$4,8 \cdot 10^{-8}$	$1,3 \cdot 10^{-11}$	$1,6 \cdot 10^{-14}$
30	$3,9 \cdot 10^{-4}$	$9,6 \cdot 10^{-9}$	$1,2 \cdot 10^{-12}$	$2,6 \cdot 10^{-14}$
40	$2,2 \cdot 10^{-4}$	$3,0 \cdot 10^{-9}$	$2,0 \cdot 10^{-13}$	$1,7 \cdot 10^{-12}$

Les erreurs relatives d'arrondi se situent au niveau de 10^{-14}, ce qui influe sensiblement sur les valeurs de la dernière colonne.

3.3. Equation à coefficients discontinus

Le paragraphe 3.1 a été consacré au problème de Dirichlet pour l'équation de diffusion à coefficients réguliers. Les équations de diffusion qu'on se propose d'étudier dans les pages suivantes, sont à coefficients discontinus. On a déjà dit au début de ce chapitre que ces problèmes jouent un rôle de premier ordre dans de nombreuses applications. D'autre part, il y a lieu de noter que l'appareil mathématique des chapitres précédents a été basé sur la propriété, pour une solution régulière, d'être développée suivant les puissances d'un paramètre petit qui a été en l'occurrence le pas du réseau. Il est naturel que ce développement n'a pas de sens en ce qui concerne les points de discontinuité des coefficients. En effet, même a dérivée première de la solution subit une discontinuité en ces points. La méthode sera donc modifiée en conséquence, et on la justifiera pour une classe plus vaste de problèmes à coefficients discontinus.

Soit le problème de Dirichlet

$$-(pu')' + qu = f, \tag{3.1}$$

$$u(0) = u_0, \qquad u(1) = u_1. \tag{3.2}$$

où

$$p(x) \geqslant c_1 > 0, \qquad q(x) \geqslant 0.$$

Les fonctions p, q, f peuvent posséder des discontinuités de première espèce en un nombre fini de points du segment [0, 1]. On simplifie les calculs si l'on se limite à un seul point de discontinuité $\xi \in (0, 1)$. On admet que l'équation (3.1) a lieu sur chacun des intervalles $(0, \xi)$, $(\xi, 1)$ et on exige au point ξ

$$u(\xi + 0) = u(\xi - 0), \tag{3.3}$$

$$p(\xi + 0)u'(\xi + 0) - p(\xi - 0)u'(\xi - 0) = g, \tag{3.4}$$

avec g une constante fixe (conditions de transmission).

On désigne par $v(\xi \pm 0)$ pour v arbitraire les expressions respectives

$$\lim_{\substack{\delta \to 0\\ \delta > 0}} v(\xi + \delta) \quad \text{et} \quad \lim_{\substack{\delta \to 0\\ \delta > 0}} v(\xi - \delta).$$

On introduit les classes fonctionnelles Q_ξ^k, avec k entier naturel. On admet que $v \in Q_\xi^k$ si elle est définie sur [0, 1] et possède des dérivées continues par morceaux jusqu'à l'ordre k, la fonction et ses dérivées ne pouvant subir d'autres discontinuités que des discontinuités de première espèce au point ξ.

Théorème 3.1. *Si les coefficients du problème* (3.1) *à* (3.4) *sont supposés tels que*

$$p \in Q_\xi^{r+1}, \qquad f, q \in Q_\xi^r, \qquad r \geqslant 2, \tag{3.5}$$

il existe une solution u unique ayant les propriétés suivantes:

$$u \in Q_\xi^{r+2}, \qquad u \in C[0, 1]. \tag{3.6}$$

Démonstration. On cherche la solution u sous forme de somme $u = w_1 + w_2$, la fonction w_1 étant définie par la formule

$$w_1(x) = \begin{cases} u_0 + ax & \text{si } x \in [0, \xi], \\ u_1 + (1 - x)\, b & \text{si } x \in [\xi, 1], \end{cases}$$

$$a = \frac{p(\xi + 0)\,(u_1 - u_0) - (1 - \xi)\, g}{\xi\, p(\xi + 0) + (1 - \xi)\, p\,(\xi - 0)},$$

$$b = \frac{p\,(\xi - 0)\,(u_0 - u_1) - \xi\, g}{\xi\, p\,(\xi + 0) + (1 - \xi)\, p\,(\xi - 0)}.$$

La fonction w_1 ainsi construite jouit des propriétés suivantes:

$$w_1(0) = u_0, \qquad w_1(1) = u_1,$$
$$w_1(\xi + 0) - w_1(\xi - 0) = 0,$$
$$p\,(\xi + 0)\, w_1'(\xi + 0) - p\,(\xi - 0)\, w_1'(\xi - 0) = g,$$
$$-(p\, w_1')' + q\, w_1 = z_1,$$

où

$$z_1(x) = \begin{cases} -ap'(x) + q(x)(u_0 + ax) & \text{si } x \in (0, \xi), \\ bp'(x) + q(x)\,(u_1 + (1 - x)\, b) & \text{si } x \in (\xi, 1). \end{cases}$$

On définit w_2 comme solution du problème avec conditions homogènes aux extrémités du segment $[0, 1]$ et au point ξ:

$$\left.\begin{aligned} & w_2(0) = 0, \quad w_2(1) = 0, \\ & w_2(\xi + 0) - w_2(\xi - 0) = 0, \\ & p\,(\xi + 0)\, w_2'(\xi + 0) - p\,(\xi - 0)\, w_2'(\xi - 0) = 0, \\ & -(p w_2')' + q w_2 = f - z_1. \end{aligned}\right\} \tag{3.7}$$

Il est connu (voir [58]) que le problème (3.7) a une solution unique continue sur $[0, 1]$. C'est cette solution qu'on assimile à w_2.

La fonction $u = w_1 + w_2$ est évidemment solution du problème (3.1) à (3.4). Cette solution est la seule solution du problème. En effet, on raisonne par l'absurde et on suppose que le problème concerné possède une autre solution, disons u_1, auquel cas la différence $u_1 - u$ vérifie le problème avec conditions homogènes (3.7), où $f - z_1 = 0$.

Or, ce problème n'a pas d'autre solution à part la solution triviale, si bien que $u \equiv u_1$.

Ainsi, u est la solution, et elle est continue sur $[0, 1]$ parce qu'il en est de même des deux termes w_1 et w_2. Il en résulte que $u(\xi)$ est une quantité finie. Soit deux problèmes aux limites

$$-(pv_1')' + qv_1 = f \quad \text{sur } (0, \xi),$$
$$v_1(0) = u_0, \qquad v_1(\xi) = u(\xi)$$

et

$$-(pv_2')' + qv_2 = f \quad \text{sur } (\xi, 1),$$
$$v_2(\xi) = u(\xi), \qquad v_2(1) = u_1.$$

Les coefficients p, q, f ne subissent pas de discontinuités dans le domaine où les équations sont définies, si bien qu'il y a existence et unicité pour les deux problèmes par le théorème 1.1, et $v_1 \in C^{r+2}[0, \xi]$, $v_2 \in C^{r+2}[\xi, 1]$. Mais la fonction u est également solution de ces problèmes. Aussi

$$u(x) = \begin{cases} v_1(x) & \text{si } x \in [0, \xi], \\ v_2(x) & \text{si } x \in [\xi, 1] \end{cases}$$

jouit des propriétés (3.6), c.q.f.d.

On pose $h_1 = \xi/N_1$, $h_2 = (1 - \xi)/N_2$ et on discrétise le domaine de façon qu'il contienne le point de discontinuité des coefficients ξ:

$$\begin{aligned} \omega_{h_1} &= \{x_i = ih_1;\ i = 1, \ldots, N_1 - 1\}, \\ \omega_{h_2} &= \{x_i = \xi + (i - N_1) h_2;\ i = N_1 + 1, \ldots, N_1 + N_2 - 1\}. \quad (3.8) \\ \omega_h &= \omega_{h_1} \cup \omega_{h_2} \cup \{x_{N_1} = \xi\}. \end{aligned}$$

On note qu'à l'intérieur de chaque région de régularité des coefficients, les pas du réseau sont réguliers. On introduit l'ensemble des points médians

$$\bar{\omega}_h = \{x_{i+1/2} = (x_i + x_{i+1})/2; \qquad i = 0, 1, \ldots, N_1 + N_2 - 1\}, \quad (3.9)$$

et on considère les équations aux différences

$$-(pu_{\bar{x}}^h)_{\hat{x}} + qu^h = f \qquad \text{sur } \omega_{h_1}, \tag{3.10}$$

$$-(pu_{\bar{x}}^h)_{\hat{x}} + qu^h = f \qquad \text{sur } \omega_{h_2}. \tag{3.11}$$

Les réseaux ω_{h_1} et ω_{h_2} étant réguliers de pas h_1 et h_2 respectivement, les dérivées aux différences ont une signification simple. Comme les fonctions q et f sont définies en ξ de façon multivoque, on mo-

difie (3.10) et (3.11) pour tenir compte de deux valeurs de chacune d'elles. On a notamment

$$-(pu^h_{\bar{x}})_{\hat{x}}\,|_{\xi} + \frac{h_1 q(\xi-0) + h_2 q(\xi+0)}{h_1+h_2}\, u^h(\xi) = \frac{h_1 f(\xi-0) + h_2 f(\xi+0)}{h_1+h_2} - \frac{2g}{h_1+h_2}. \qquad (3.12)$$

Conformément aux notations principales pour les réseaux non réguliers,

$$(pu^h_{\bar{x}})_{\hat{x}}\,|_{\xi} = (pu^h_{\bar{x}})_{\hat{x}}\,|_{x_{N_1}} = \frac{(pu^h_{\bar{x}})\,|_{x_{N_1+1/2}} - (pu^h_{\bar{x}})\,|_{x_{N_1-1/2}}}{x_{N_1+1/2} - x_{N_1-1/2}} =$$

$$= \frac{p\left(\xi + \frac{h_2}{2}\right)\left(u^h(\xi+h_2) - u^h(\xi)\right)}{h_2(h_1+h_2)/2} - \frac{p\left(\xi - \frac{h_1}{2}\right)\left(u^h(\xi) - u^h(\xi-h_1)\right)}{h_1(h_1+h_2)/2}.$$

Avec les conditions aux limites

$$u^h(0) = u_0, \qquad u^h(1) = u_1, \qquad (3.13)$$

le système d'équations aux différences se trouve décrit. Il y a existence, unicité et stabilité pour le schéma construit (voir [41]).

Un lecteur désireux de se familiariser avec les principes de construction des équations de la forme (3.12) est prié de consulter [43], [65], [112] qui exposent, en particulier, les procédés les plus répandus pour bâtir les équations aux différences à coefficients discontinus.

On note qu'on n'extrapole avec succès qu'en présence d'un seul paramètre de discrétisation. Or, notre schéma en possède deux. On se débarrasse du degré de liberté superflu en imposant aux pas de vérifier la relation

$$h_1 = c_2 h_2, \qquad (3.14)$$

où c_2 ne varie pas avec h_1 et h_2, ce qui permet d'introduire un seul paramètre ($\hbar = \sqrt{h_1 h_2}$) caractéristique du domaine discrétisé:

$$h_1 = \sqrt{c_2}\,\hbar, \qquad h_2 = \hbar/\sqrt{c_2}. \qquad (3.15)$$

THÉORÈME 3.2. *On suppose que les coefficients du problème différentiel* (3.1) *à* (3.4) *vérifient les conditions de régularité* (3.5) *et qu'on est dans l'hypothèse* (3.14) *pour le problème approché* (3.10) *à* (3.13). *La solution approchée admet le développement*

$$u^h = u + \sum_{j=1}^{l} \hbar^{2j} v_j + \hbar^r \eta^h \quad sur \quad \bar{\omega}_h. \qquad (3.16)$$

Ici $l = [(r-1)/2]$, *les fonctions* v_j *sont dans* Q_ξ^{r+2-2j} *et ne dépendent pas de* $\hbar$, *la fonction discrète* $\eta^\hbar$ *est de module uniformément borné sur* $\bar{\omega}_\hbar$.

Démonstration. On pose

$$v_0 = u,$$

et soit la famille de problèmes différentiels

$$-(pv_j)' + qv_j =$$

$$= \sum_{1 \leqslant s+k \leqslant j} \frac{\left(pv_{j-s-k}^{(2s+1)}\right)^{(2k+1)} c_2^{s+k}}{2^{2s+2k}(2s+1)!\,(2k+1)!} \quad \text{sur} \quad (0,\ \xi), \tag{3.17}$$

$$-(pv_j')' + qv_j = \sum_{1 \leqslant s+k \leqslant j} \frac{pv\left(_{j-s-k}^{(2s+1)}\right)^{(2k+1)} c_2^{-s-k}}{2^{2s+2k}(2s+1)!\,(2k+1)!} \text{ sur } (\xi,\ 1), \tag{3.18}$$

$$v_j(0) = 0, \qquad v_j(1) = 0, \tag{3.19}$$

$$v_j(\xi+0) - v_j(\xi-0) = 0, \tag{3.20}$$

$$p(\xi+0)\, v_j'(\xi+0) - p(\xi-0)\, v_j'(\xi-0) =$$

$$= -\sum_{1 \leqslant s+k \leqslant j} \frac{c_2^{-s-k}\left(pv_{j-s-k}^{(2k+1)}\right)^{(2s)}\Big|_{\xi+0} - c_2^{s+k}\left(pv_{j-s-k}^{(2k+1)}\right)^{(2s)}\Big|_{\xi-0}}{(2s)!\,(2k+1)!\,2^{2s+2k}} \tag{3.21}$$

$$j = 1, \ldots, l.$$

On suppose que $v_0, \ldots, v_{j-1}$ sont connues et que $v_k \in Q_\xi^{r+2-2k}$, $0 \leqslant k \leqslant j-1$. Dans ce cas, on assimile les égalités (3.17) à (3.21) au problème de déterminer v_j. Si l'on prolonge les valeurs des seconds membres de (3.17), (3.18) par les limites à gauche et à droite en ξ, elles deviennent des fonctions $r-2j$ fois continûment dérivables sur $[0,\ \xi]$ et $[\xi,\ 1]$ respectivement. Du moment que les constantes du second membre de (3.21) sont finies en tant que limites des fonctions ayant une discontinuité de première espèce, la fonction v_j est définie de façon unique et appartient à Q_ξ^{r-2j+2} (théorème 3.1).

On démontre, pour la fonction discrète

$$\eta^\hbar = \hbar^{-r}\left(u^\hbar - \sum_{j=1}^{l} \hbar^{2j} v_j\right) \qquad \text{sur } \bar{\omega}_\hbar. \tag{3.22}$$

la propriété d'être bornée lorsque $\hbar \to 0$. On résout l'équation (3.22) par rapport à $u^{\hbar}$ et on porte le résultat dans (3.10), il vient

$$-\sum_{j=0}^{l} \hbar^{2j} (p (v_j)_{\bar{x}})_x - \hbar^r (p \eta^{\hbar}_{\bar{x}})_x + \sum_{j=0}^{l} \hbar^{2j} q v_j + \hbar^r q \eta^{\hbar} = f. \quad (3.23)$$

On applique aux termes $(p (v_j)_{\bar{x}})_x$ le lemme 1.2, § 7.1 :

$$(p (v_j)_{\bar{x}})_x =$$

$$= \sum_{k=0}^{l-j} h_1^{2k} \frac{2^{-2k}}{(2k+1)!} \sum_{s=0}^{l-j-k} h_1^{2s} \frac{2^{-2s}}{(2s+1)!} (p v_j^{(2k+1)})^{(2s+1)} +$$
$$+ h_1^{r-2j} \varkappa_j^{\hbar}, \quad (3.24)$$

où

$$| \varkappa_j^{\hbar} (x) | \leqslant c_3 \qquad \forall\, x \in \omega_{h_1}.$$

On porte le développement obtenu dans (3.23). On réduit les termes semblables (on se rappelle que $h_1 = \sqrt{c_2}\, \hbar$) et on aboutit à l'égalité

$$\sum_{j=0}^{l} \hbar^{2j} \left(- \sum_{0 \leqslant k+s \leqslant j} \frac{c_2^{k+s} \left(p v_{j-s-k}^{(2k+1)}\right)^{(2s+1)}}{(2k+1)!(2s+1)!\, 2^{2k+2s}} + q v_j \right) -$$
$$- \hbar^r (p \eta^{\hbar}_{\bar{x}})_x - \hbar^r \sum_{j=0}^{l} c_2^{r/2-j} \varkappa_j^{\hbar} + \hbar^r q \eta^{\hbar} = f.$$

Avec l'équation (3.1) juste pour v_0 et le procédé de définition des fonctions v_j, on réduit tous les termes de l'ordre de $\hbar^0, \ldots, \hbar^{2l}$. On divise les termes restants par $\hbar^r$ et on a finalement

$$-(p \eta^{\hbar}_{\bar{x}})_x + q \eta^{\hbar} = \sum_{j=0}^{l} c_2^{r/2-j} \varkappa_j^{\hbar} \quad \text{sur} \quad \omega_{h_1}, \quad (3.25)$$

où

$$\max_{x \in \omega_{h_1}} \sum_{j=0}^{l} c_2^{r/2-j} | \varkappa_j^{\hbar} (x) | \leqslant c_4 = c_3 \max \{1, c_2^{r/2}\}. \quad (3.26)$$

S'agissant du réseau ω_{h_2}, on trouve de même

$$-(p \eta^{\hbar}_{\bar{x}})_x + q \eta^{\hbar} = - \sum_{j=0}^{l} c_2^{-r/2+j} \varkappa_j^{\hbar} \quad \text{sur} \quad \omega_{h_2}. \quad (3.27)$$

où

$$\max_{x\in\omega_{h_2}} \sum_{j=0}^{l} c_2^{j-r/2} \,|\, \varkappa_j^h(x) \,|\, \leqslant c_5 = c_3 \max\{1,\ c_2^{-r/2}\}. \qquad (3.28)$$

On substitue l'expression de u^h dans l'équation associée au point ξ :

$$-\sum_{j=0}^{l} \hbar^{2j}(p\,(v_j)_{\bar{x}})_{\hat{x}} - \hbar^r(p\eta^h_{\bar{x}})_{\hat{x}} +$$

$$+\sum_{j=0}^{l} \frac{h_1 q\,(\xi-0)+h_2 q\,(\xi+0)}{h_1+h_2}\,\hbar^{2j}\,v_j + \hbar^r\,\frac{h_1 q\,(\xi-0)+h_2 q\,(\xi+0)}{h_1+h_2)}\,\eta_l^h =$$

$$= \frac{h_1 f\,(\xi-0)+h_2 f\,(\xi+0)}{h_1+h_2} - \frac{2g}{h_1+h_2}. \qquad (3.29)$$

On applique le lemme 1.1, § 7.1 à la différence divisée $(v_j)_{\bar{x}}$:

$$\left(v_j\left(\xi+\frac{h_2}{2}\right)\right)_{\hat{x}} = \sum_{k=0}^{l-j} h_2^{2k}\,\frac{2^{-2k}}{(2k+1)\,!}\,v^{(2k+1)}\left(\xi+\frac{h_2}{2}\right) + h_2^{r-2j+1}\,\sigma_j,$$

où

$$|\,\sigma_j\,| \leqslant c_6.$$

On multiplie par $p\,(\xi+h_2/2)$ et on développe chaque terme $pv_j^{(2k+1)}$ en formule de Taylor sans oublier que les dérivées correspondantes admettent en ξ une limite finie à droite :

$$p\,(v_j)_{\bar{x}}\,|_{\xi+h_2/2} =$$

$$= \sum_{k=0}^{l-j} h_2^{2k}\,\frac{2^{-2k}}{(2k+1)\,!}\sum_{s=0}^{r-2j-2k} h_2^s\,\frac{2^{-s}}{s\,!}\,(pv_j^{(2k+1)})^{(s)}|_{\xi+0} + h_2^{r-2j+1}\,\rho_j^+\,, \qquad (3.30)$$

où

$$|\,\rho_j^+\,| \leqslant c_7, \qquad j = 0,\ 1,\ \ldots,\ l. \qquad (3.31)$$

On a de même pour la limite à gauche :

$$p\,(v_j)_{\bar{x}}\,|_{\xi-h_1/2} =$$

$$= \sum_{k=0}^{l-j} h_1^{2k}\,\frac{2^{-2k}}{(2k+1)\,!}\sum_{s=0}^{r-2j-2k} (-h_1)^s\,\frac{2^{-s}}{s\,!}\,(pv_j^{(2k+1)})^{(s)}|_{\xi-0} + h_1^{r-2j+1}\,\rho_j^-\,. \qquad (3.32)$$

où

$$|\,\rho_j^-\,| \leqslant c_8, \qquad j = 0,\ 1,\ \ldots,\ l. \qquad (3.33)$$

On fait la différence de (3.30) et (3.32) et on divise membre à membre par $(h_1 + h_2)/2$:

$$(p\,(v_j)_{\bar{x}})_{\hat{x}}|_\xi =$$

$$= \sum_{k=0}^{l-j} \sum_{s=0}^{l-j-k} \frac{2^{-2s-2k}}{(2s+1)!\,(2k+1)!} \left[\frac{h_2^{2k+2s+1}}{h_1+h_2} \left(pv_j^{(2k+1)}\right)^{(2s+1)} |_{\xi+0} + \right.$$

$$\left. + \frac{h_1^{2k+2s+1}}{h_1+h_2} \left(pv_j^{(2k+1)}\right)^{(2s+1)} |_{\xi-0} \right] +$$

$$+ \sum_{k=0}^{l-j} \sum_{s=0}^{l-j-k} \frac{2^{-2s-2k+1}}{(2s)!\,(2k+1)!} \left[\frac{h_2^{2k+2s}}{h_1+h_2} \left(pv_j^{(2k+1)}\right)^{(2s)} |_{\xi+0} - \right.$$

$$\left. - \frac{h_1^{2k+2s}}{h_1+h_2} \left(pv_j^{(2k+1)}\right)^{(2s)} |_{\xi-0} \right] + \frac{\hbar^{r-2j}}{h_1+h_2} \mu_j, \qquad (3.34)$$

où

$$|\mu_j| \leqslant c_9, \qquad j = 0,\ 1,\ \ldots,\ l, \qquad (3.35)$$

par suite des estimations (3.31), (3.33).

On a partagé les termes du second membre de (3.34) en deux groupes dont l'un est formé de termes qui subsistent sur la partie régulière de $\omega_\hbar$ et l'autre de ceux qui s'y réduisent. On porte les développements obtenus dans (3.29), il vient compte tenu des relations (3.15) entre h_i et $\hbar$:

$$-\sum_{j=0}^{l} \hbar^{2j} \left(\sum_{0 \leqslant k+s \leqslant j} \frac{2^{-2s-2k}}{(2s+1)!\,(2k+1)!} \left[\frac{h_2 c_2^{-s-k} \left(pv_{j-s-k}^{(2k+1)}\right)^{(2s+1)} |_{\xi+0}}{h_1+h_2} + \right.\right.$$

$$\left.\left. + \frac{h_1 c_2^{k+s} \left(pv_{j-s-k}^{(2k+1)}\right)^{(2s+1)} |_{\xi-0}}{h_1+h_2} \right] + \frac{h_2 q\,(\xi+0) + h_1 q\,(\xi-0)}{h_1+h_2} v_j\,(\xi) \right) -$$

$$- \sum_{j=0}^{l} \hbar^{2j} \sum_{0 \leqslant k+s \leqslant j} \frac{2^{-2s-2k+1}}{(2s)!\,(2k+1)!} \times$$

$$\times \left[\frac{c_2^{-k-s} \left(pv_{j-s-k}^{(2k+1)}\right)^{(2s)} |_{\xi+0}}{h_1+h_2} - \frac{c_2^{k+s} \left(pv_{j-s-k}^{(2k+1)}\right)^{(2s)} |_{\xi-0}}{h_1+h_2} \right] -$$

$$- \hbar^r (p\,\eta^\hbar_{\bar{x}})_{\hat{x}} |_\xi + \hbar^r \frac{h_1 q(\xi-0) + h_2 q\,(\xi+0)}{h_1+h_2} \eta^\hbar\,(\xi) +$$

$$+ \frac{\hbar^r}{h_1+h_2} \nu^\hbar = \frac{h_1 f\,(\xi-0) + h_2 f\,(\xi+0)}{h_1+h_2} - \frac{2g}{h_1+h_2}, \qquad (3.36)$$

et les estimations (3.35) entraînent pour $\nu^\hbar$:

$$|\nu^\hbar| \leqslant c_{10}. \qquad (3.37)$$

On simplifie l'identité (3.36) si l'on se rappelle que $v_0 = u$. Aussi (3.1) entraîne les égalités

$$-(pv_0')'\,|_{\xi+0} + q(\xi+0)\,v_0(\xi) = f(\xi+0),$$
$$-(pv_0')'\,|_{\xi-0} + q(\xi-0)\,v_0(\xi) = f(\xi-0).$$

En combinant celles-ci, on a

$$-\frac{h_1(pv_0')'\,|_{\xi-0} + h_2(pv_0')'\,|_{\xi+0}}{h_1+h_2} + \frac{h_1q(\xi-0)+h_2q(\xi+0)}{h_1+h_2}\,v_0(\xi) = \\ = \frac{h_1f(\xi-0)+h_2f(\xi+0)}{h_1+h_2}. \quad (3.38)$$

On divise les deux membres de (3.4) par $(h_1+h_2)/2$:

$$\frac{2}{h_1+h_2}\left((pv_0')\,|_{\xi+0} - (pv_0')\,|_{\xi-0}\right) = \frac{2g}{h_1+h_2}. \quad (3.39)$$

Les égalités (3.38) et (3.39) aidant, on réduit dans (3.36) tous les termes du second membre et les termes à gauche qui sont associés à l'indice $j = 0$.

Avec des raisonnements analogues, on tire de (3.17) et (3.18)

$$-\sum_{0<s+k<j} \frac{h_1c_2^{\,k+s}\left(pv_{j-s-k}^{(2k+1)}\right)^{(2s+1)}\Big|_{\xi-0} + h_2c_2^{\,-k-s}\left(pv_{j-s-k}^{(2k+1)}\right)^{(2s+1)}\Big|_{\xi+0}}{(h_1+h_2)\,2^{2s+2k}\,(2s+1)!\,(2k+1)!} + \\ + \frac{h_1q(\xi-0)+h_2q(\xi+0)}{h_1+h_2}\,v_j(\xi) = 0. \quad (3.40)$$

On porte tous les termes de la condition de transmission (3.21) dans le premier membre et on divise par $(h_1+h_2)/2$, il vient

$$\sum_{0<k+s<j} \frac{c_2^{\,-k-s}\left(pv_{j-s-k}^{(2k+1)}\right)^{(2s)}\Big|_{\xi+0} - c_2^{\,k+s}\left(pv_{j-s-k}^{2k+1)}\right)^{(2s)}\Big|_{\xi-0}}{(h_1+h_2)\,(2s)!\,(2k+1)!\,2^{2s+2k-1}} = 0. \quad (3.41)$$

Ainsi, les égalités (3.40) et (3.41) permettent de réduire beaucoup de termes de (3.36). On divise les termes restants par $\hbar^r$:

$$-(p\eta_\xi^{\hbar})_\xi\,|_\xi + \frac{h_1q(\xi-0)+h_2q(\xi+0)}{h_1+h_2}\,\eta^{\hbar}(\xi) = -\frac{\nu^{\hbar}}{h_1+h_2}. \quad (3.42)$$

La disparition des termes d'ordre inférieur à $\hbar^2$ est parfaitement normale car nous avons choisi les conditions (3.17), (3.18), (3.21) dans l'intention d'y parvenir.

On évalue la solution η^h à partir de (3.25), (3.27), (3.42) si l'on tient compte des relations

$$\begin{aligned} \eta^h(0) &= h^{-r}\left(u^h(0) - \sum_{j=0}^{l} h^{2j} v_j(0)\right) = 0, \\ \eta^h(1) &= h^{-r}\left(u^h(1) - \sum_{j=0}^{l} h^{2j} v_j(1)\right) = 0, \end{aligned} \tag{3.43}$$

ayant lieu par suite des conditions aux limites (3.2), (3.13) et (3.19). En effet, il existe pour toute fonction égale à 0 aux extrémités du segment [0, 1] telle que

$$-(p\eta^h_{\bar{x}})_{\hat{x}} + q\eta^h = w \quad \text{sur} \quad \omega_{h_1} \cup \omega_{h_2},$$

$$-(p\eta^h_{\bar{x}})_{\hat{x}}|_\xi + \frac{h_1 q(\xi - 0) + h_2 q(\xi + 0)}{h_1 + h_2}\eta^h(\xi) = w(\xi),$$

une fonction de Green discrète $G(x, t)$ pour laquelle (voir [67])

$$\eta^h(x) = \sum_{t \in \omega_{h_1}} G(x, t)\, w(t)\, h_1 + G(x, \xi)\, w(\xi)\frac{h_1 + h_2}{2} + \sum_{t \in \omega_{h_2}} G(x, t)\, w(t)\, h_2.$$

La fonction de Green est de plus uniformément bornée:

$$|G(x, t)| \leqslant \frac{1}{c_1} \qquad \forall\, x \in \bar{\omega}_h, \qquad \forall\, t \in \omega_h.$$

Aussi

$$|\eta^h(x)| \leqslant \frac{1}{c_1}\left(\sum_{t \in \omega_{h_1}} |w(t)|\, h_1 + |w(\xi)|\,\frac{h_1 + h_2}{2} + \sum_{t \in \omega_{h_2}} |w(t)\, h_2|\right).$$

On utilise, pour évaluer les valeurs de $w(t)$, les inégalités (3.26), (3.28), (3.37) et on est conduit aux relations

$$|\eta^h(x)| \leqslant \frac{1}{c_1}\left(\sum_{t \in \omega_{h_1}} c_4 h_1 + \frac{|\nu^h|}{h_1 + h_2}\,\frac{h_1 + h_2}{2} + \sum_{t \in \omega_{h_2}} c_5 h_2\right) \leqslant$$

$$\leqslant \frac{1}{c_1}\left(c_4\,\xi + \frac{c_{10}}{2} + c_5(1 - \xi)\right) = c_{11}.$$

Etant donné l'arbitraire laissé sur le point $x \in \bar{\omega}_h$, on vient de démontrer l'estimation

$$\max_{x \in \bar{\omega}_h} |\eta^h(x)| \leqslant c_{11} \tag{3.44}$$

qui justifie le développement (3.16). Le théorème 3.2 se trouve démontré.

Avec le développement (3.16), on peut extrapoler moyennant $l+1$ solutions du problème aux différences (3.10) à (3.13) qui sont associées aux réseaux de h_i différents; cela étant, la précision de la solution corrigée est $O(h^r)$. On note que l'extrapolation sera en défaut si l'on prend une suite de réseaux tels que le rapport N_1/N_2 varie. En effet, on a par suite de la condition $h_1 = c_2 h_2$ et des rapports $h_1 = \xi/N_1$, $h_2 = (1-\xi)/N_2$:

$$N_1/N_2 = \frac{\xi}{c_2(1-\xi)}.$$

Ainsi, le rapport N_1/N_2 défini par le réseau initial doit être le même pour tous les réseaux suivants.

Prenons donc $N_1 = K$ et $N_2 = L$ pour valeurs primitives définissant le réseau ω_{h_1} et formons les réseaux ω_{h_i}, avec $N_1 = iK$ et $N_2 = iL$ entiers. On résout sur chaque ω_{h_i} le problème aux différences (3.10) à (3.13). Soit u^{h_i} les solutions obtenues. Toutes ces fonctions sont définies sur $\bar{\omega}_{h_1}$. Formons la combinaison linéaire

$$U = \sum_{k=1}^{l+1} \gamma_k u^{h_k}, \tag{3.45}$$

où les poids

$$\gamma_k = \frac{2(-1)^{l+k+1} k^{2l+2}}{(l-k+1)!(l+k+1)!}. \tag{3.46}$$

Théorème 3.3. *Hypothèses du théorème 3.2. La solution corrigée* (3.45) *admet l'estimation*

$$\max_{x \in \bar{\omega}_{h_1}} |U(x) - u(x)| \leq h_1^r c_{12}, \tag{3.47}$$

avec la constante indépendante de h_i.

Démonstration. Le théorème 3.2 implique les développements

$$u^{h_k} = u + \sum_{j=1}^{l} h_k^{2j} v_j + h_k^r \eta^{h_k} \quad \text{sur} \quad \bar{\omega}_{h_1},$$

avec v_j indépendantes de k. Aussi

$$U = \sum_{k=1}^{l+1} \gamma_k u + \left(\sum_{j=1}^{l} \sum_{k=1}^{l+1} \gamma_k h_k^{2j}\right) v_j + \sum_{k=1}^{l+1} \gamma_k h_k^r \eta^{h_k}. \tag{3.48}$$

Etant donné le lemme 2.2, § 7.2, les poids γ_k vérifient les égalités

$$\sum_{k=1}^{l+1} \gamma_k = 1,$$
$$\sum_{k=1}^{l+1} \gamma_k \frac{1}{k^{2j}} = 0, \qquad j = 1, \ldots, l. \tag{3.49}$$

On multiplie la dernière égalité par $\hbar_1^{2j}$ et on a

$$\sum_{k=1}^{l+1} \gamma_k \hbar_k^{2j} = 0 \tag{3.50}$$

vue que

$$\frac{\hbar_1}{k} = \frac{\sqrt{\xi(1-\xi)}}{k\sqrt{KL}} = \frac{\sqrt{\xi(1-\xi)}}{\sqrt{kKkL}} = \hbar_k .$$

On transforme (3.48) au moyen de (3.49), (3.30):

$$U = u + \sum_{k=1}^{l+1} \gamma_k \hbar_k^r \eta^{\hbar_k} .$$

D'où

$$|U(x) - u(x)| \leqslant \sum_{k=1}^{l+1} |\gamma_k| \, |\eta^{\hbar_k}| \, \hbar_k^r, \qquad x \in \bar{\omega}_{\hbar_1} .$$

L'estimation (3.44) et γ_k sous forme explicite donnent l'inégalité (3.47), où

$$c_{12} = c_{11} \sum_{k=1}^{l+1} \frac{2\,k^{2l+2-r}}{(l-k+1)!\,(l+k+1)!} ,$$

i.e. la démonstration de théorème 3.3 se trouve achevée.

Dans le cas où l'on demandait la valeur de la fonction en un point z qui ne coïncide avec aucun nœud d'un réseau $\omega_{\hbar_i}$ quelconque, on recourait à l'interpolation. La précision $O(\hbar_i^r)$ serait atteinte avec le polynôme d'interpolation de Lagrange basé sur r points. Dans le § 2.1, nous avons pris r points les plus proches de z pour obtenir la meilleure précision espérée (voir [67]). Etant donnée la régularité par morceaux, ce procédé est inopérant dans notre cas. On doit le modifier de façon qu'on interpole sur les points d'un même domaine de régularité. Aussi on interpole pour $z \in (0, \xi)$ sur r points le plus proches qui sont éléments de l'ensemble $\omega_{\hbar_1} \cup$

$\cup \{0\} \cup \{\xi\}$, et si $z \in (\xi, 1)$, on base le polynôme d'interpolation de Lagrange sur r points le plus proches appartenant à $\omega_{h_2} \cup \{\xi\} \cup \cup \{1\}$. C'est la seule différence sérieuse avec les affirmations du § 2.1.

Si les points de discontinuité sont deux ou plus, l'algorithme diffère de celui décrit par la construction des réseaux successifs. Soit

$$0 = z_0 < z_1 < \ldots < z_{m-1} < z_m = 1$$

l'ensemble des points de discontinuité de première espèce des fonctions p, q, f et de leurs dérivées. On prend m entiers $N_i \geqslant 2$ et on construit le réseau de discrétisation $\bar{\omega}_{h_1}$ régulier à l'intérieur de chaque domaine de régularité (z_{i-1}, z_i). A cet effet, on partage chaque intervalle (z_{i-1}, z) en N_i intervalles partiels de même longueur $(z_i - z_{i-1})/N_i$ et on prend pour nœuds de $\bar{\omega}_{h_1}$ tous les points de partage et les points $z_0, \ldots, z_m$ mêmes. On procède de même en ce qui concerne les réseaux suivants à la différence que les sous-intervalles sont pour chaque $\bar{\omega}_{h_k}$ au nombre de kN_i. Le paramètre

$$h_k = \frac{1}{k\,(N_1 \ldots N_m)^{1/m}}\,.$$

Dans la suite, on opère comme pour (3.45), (3.46).

3.4. Problème de Sturm-Liouville

En physique mathématique, on voit souvent intervenir le problème de valeurs propres dont le cas le plus simple porte le nom de Sturm-Liouville. Ce problème donne un jeu de fonctions propres et de valeurs propres associées. Dans la plupart des cas pratiques intéressants, sa résolution analytique s'avère délicate, voire impossible, si bien qu'on le résout efficacement par des méthodes numériques. Or, ces méthodes introduisent toujours des erreurs d'approximation plus ou moins sérieuses, et si l'on a de règle les premières fonctions et valeurs propres avec des erreurs négligeables, les erreurs sur les éléments suivants sont d'ordinaire importantes. S'agissant d'une vaste gamme de fonctions et valeurs propres, les solutions approchées doivent donc être améliorées, et on utilise à cet effet l'extrapolation de Richardson. On part d'une famille de solutions approchées relatives à divers réseaux et on en forme une combinaison linéaire à coefficients donnés. La solution ainsi obtenue est d'ordre de précision élevé aussi bien pour les fonctions propres que pour les valeurs propres. On trouvera dans ce paragraphe une justification de l'algorithme.

Le problème de Sturm-Liouville le plus simple consiste à chercher des valeurs du paramètre λ (les valeurs propres) telles qu'il existe des solutions non triviales (les fonctions propres) de l'équation

$$-(py')' + qy = \lambda y,$$
$$p(x) \geqslant c_1 > 0, \qquad q(x) \geqslant 0, \qquad x \in (0, 1), \tag{4.1}$$

avec les conditions aux limites homogènes

$$y(0) = y(1) = 0. \tag{4.2}$$

On exige que les coefficients du problème vérifient la condition de régularité

$$p \in C^{r+1}[0, 1], \qquad q \in C^r[0, 1], \tag{4.3}$$

avec r un entier naturel.

L'équation (4.1) définissant la fonction à un facteur constant près, on entendra par fonction propre normée une fonction satisfaisant à la condition de normalisation

$$\int_0^1 (y(x))^2\,dx = 1 \tag{4.4}$$

et à

$$y'(0) > 0. \tag{4.5}$$

Citons certaines propriétés des valeurs et fonctions propres du problème (4.1) à (4.5) (voir [43], [78], [92]).

On trouve un ensemble dénombrable de valeurs propres $0 < \lambda_1 < \ldots < \lambda_n < \ldots$ associées aux fonctions propres $y_1(x), \ldots$ $\ldots, y_n(x), \ldots$, chaque valeur propre correspondant à une, et une seule, fonction propre. Les fonctions propres forment un système orthonormé, i.e.

$$\int_0^1 y_n(x)\,y_m(x)\,dx = 0$$

pour $n \neq m$. Les valeurs propres sont encadrées comme suit:

$$c_2 n^2 - c_3 \leqslant \lambda_n \leqslant c_2 n^2 + c_3, \tag{4.6}$$

avec c_2, c_3 indépendantes de n.

Quant aux fonctions propres et à leurs dérivées, on a les estimations

$$|y_n(x)| \leqslant c_4, \qquad |y_n'(x)| \leqslant c_4 n, \tag{4.7}$$

c_4 étant une constante indépendante de n et x. Ces inégalités et l'équation (4.1) permettent d'évaluer les dérivées secondes en va-

leur absolue. Les estimations des dérivées d'ordre supérieur découlent des égalités obtenues par dérivation de (4.1). Aussi

$$\max_{[0,1]} |y_n^{(k)}| \leq c_5 n^k, \qquad k = 0, 1, \ldots, r+2, \tag{4.8}$$

où la constante c_5 ne dépend pas de n, k, x.

Comment s'énonce l'analogue aux différences du problème de valeurs propres? Conformément aux notations principales, on introduit le réseau régulier $\bar{\omega}_h$, on forme le système d'équations aux différences

$$-(p y_{\bar{x}}^h)_x + q y^h = \lambda^h y^h, \qquad x \in \omega_h, \tag{4.9}$$

et on adjoint les conditions aux limites

$$y^h(0) = y^h(1) = 0. \tag{4.10}$$

Le problème de Sturm-Liouville aux différences est alors de chercher les valeurs du paramètre λ^h (les valeurs propres discrètes) associées aux solutions non nulles (les fonctions propres discrètes) du système d'équations (4.9), (4.10).

On élimine les valeurs $y^h(0)$, $y^h(1)$ entre les équations (4.9) et on introduit la notation

$$A = \begin{bmatrix} b_1 & c_1 & & & \mathbf{0} \\ a_2 & b_2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \\ \mathbf{0} & & & & c_{N-2} \\ & & & a_{N-1} & b_{N-1} \end{bmatrix},$$

$$Y^h = \begin{bmatrix} y^h(x_1) \\ y^h(x_2) \\ \vdots \\ y^h(x_{N-1}) \end{bmatrix},$$

où

$$a_i = c_{i-1} = -\frac{p(x_{i-1/2})}{h^2}, \qquad b_i = -a_i - c_i + q(x_i).$$

Le système obtenu se ramène au problème spectral algébrique usuel

$$AY^h = \lambda^h Y^h \tag{4.11}$$

de matrice A symétrique tridiagonale.

Aussi les fonctions propres discrètes sont définies à une constante multiplicative près. Pour en dégager l'une quelconque, on fait la convention qu'elles vérifient la condition de normalisation

$$\sum_{x\in\omega_h} (y^h(x))^2 h = 1 \tag{4.12}$$

et

$$y^h(x_1) > 0. \tag{4.13}$$

Ci-dessous plusieurs propriétés du problème de Sturm-Liouville aux différences (voir [43], [83]).

Il existe $N-1$ valeurs propres discrètes $0 < \lambda_1^h < \ldots < \lambda_{N-1}^h$ et $N-1$ fonctions propres discrètes associées $y_1^h(x), \ldots, y_{N-1}^h(x)$ définies sur ω_h. A chaque valeur propre il correspond une seule fonction propre. Les fonctions propres $\{y_n(x)\}$ constituent le système orthonormé

$$\sum_{x\in\omega_h} y_i^h(x)\, y_j^h(x)\, h = \delta_{ij}.$$

δ_{ij} étant le symbole de Kronecker.

Les valeurs propres discrètes sont encadrées comme suit :

$$c_6 n^2 \leqslant \lambda_n^h \leqslant c_7 n^2, \qquad 1 \leqslant n \leqslant N-1, \tag{4.14}$$

où c_6, c_7 sont indépendantes de n et h.

Si l'on est dans la condition de régularité (4.3), alors

$$\max_{x\in\omega_h} |y_n^h(x)| \leqslant c_8 n^{1/2}. \tag{4.15}$$

la constante c_8 étant indépendante de n et h.

On fixe n et on développe suivant les puissances de h:

$$y_n^h(x) = \sum_{s=1}^{l} h^{2s} v_s(x) + h^r \eta^h(x), \qquad x \in \bar\omega_h. \tag{4.16}$$

$$\lambda_n^h = \sum_{s=0}^{l} h^{2s} \sigma_s + h^r \rho^h, \tag{4.17}$$

où $l = [(r-1)/2]$; les fonctions $v_s(x)$ et les constantes σ_s ne dépendent pas de h; ρ^h et la fonction discrète η^h sont uniformément bornées par des constantes indépendantes de h.

On suppose que $v_j \in C^{r-2j+2}[0, 1]$ et on substitue (4.16), (4.17) dans l'égalité (4.9). Étant donnée la relation

$$-(p(v_j)_{\bar{x}})_{\hat{x}} =$$

$$= -\sum_{0 \leqslant k+s \leqslant l-j} h^{2k+2s} \frac{(pv_j^{(2s)+1})^{(2k+1)}}{2^{2k+2s}(2s+1)!(2k+1)!} + h^{r-2j}\mu_j^h \text{ sur } \omega_h. \quad (4.18)$$

où

$$\max_{x \in \omega_h} |\mu_j^h(x)| \leqslant c_{11}, \quad (4.19)$$

on obtient par réduction des termes semblables

$$\sum_{j=0}^{l} h^{2j}\left(-\sum_{0 \leqslant k+s \leqslant j} \frac{(pv_{j-k-s}^{(2k+1)})^{(2s+1)}}{(2k+1)!(2s+1)2!^{2k+2s}} + qv_j\right) -$$

$$- h^r(p\eta_{\bar{x}}^h)_{\hat{x}} - h^r\sum_{j=0}^{l}\mu_j^h + h^r q\eta^h =$$

$$= \sum_{j=0}^{l} h^{2j}\sum_{k=0}^{j}\sigma_k v_{j-k} + \sum_{j=l+1}^{2l} h^{2j}\sum_{k=j-l}^{l}\sigma_k v_{j-k} +$$

$$+ h^r\lambda_n^h\eta^h + h^r\rho^h\sum_{j=0}^{l} h^{2j} v_j. \quad (4.20)$$

Ici on a introduit les notations $\sigma_0 = \lambda_n$, $v_0 = y_n$.

Dans les paragraphes précédents, notre tactique était la suivante. Avec des fonctions v_j choisies convenablement, on supprimait tous les termes d'ordre inférieur. On divisait les termes restants par h^r et on obtenait une équation discrète en η^h dont le second membre contenait exclusivement des quantités bornées. On évaluait η^h à l'aide de la propriété de stabilité du problème aux différences.

Nous procéderons de même en ce qui concerne le problème spectral. On commence par établir des conditions suffisantes sous lesquelles les constantes σ_j et les fonctions v_j se trouvent définies. On identifie les coefficients de h^0 dans les deux membres de (4.20), il vient

$$-(pv_0')' + qv_0 = \sigma_0 v_0. \quad (4.21)$$

On note qu'il suffit de prendre $\sigma_0 = \lambda_n$ et $v_0 = y_n$ pour que l'égalité ait lieu en tous les points de ω_h.

On identifie les coefficients de h^2:

$$-(pv_1')' + qv_1 - \sum_{k+s=1} \frac{(pv_0^{(2k+1)})^{(2s+1)}}{(2k+1)!(2s+1)!4} = \sigma_0 v_1 + \sigma_1 v_0.$$

Cette relation permet de définir v_1 et σ_1. Mettons-la sous la forme

$$-(pv_1')' + qv_1 - \sigma_0 v_1 = \frac{(pv_0')'''}{24} + \frac{(pv_0''')'}{24} + \sigma_1 v_0 \qquad (4.22)$$

et ajoutons les conditions aux limites

$$v_1(0) = v_1(1) = 0. \qquad (4.23)$$

Ces conditions découlent de l'hypothèse sur l'existence du développement (4.16). En effet, on obtient par identification des coefficients des puissances paires de h de (4.16), $x = 0$, $x = 1$, les relations

$$v_j(0) = v_j(1) = 0, \qquad j = 0, 1, \ldots, l, \qquad (4.24)$$

dont la première est automatiquement satisfaite pour $y_0 = y_n$.

On observe que l'équation (4.22) est prolongeable au segment $[0, 1]$ tout entier, mais que le problème correspondant est à opérateur dégénéré parce que $\sigma_0 = \lambda_n$ est valeur propre.

Lemme 4.1. *Soit $f \in C^k$ $[0, 1]$ $(k \leqslant l)$, λ_n une valeur propre du problème* (4.1) *à* (4.5) *et y_n la fonction propre associée. Le problème*

$$\begin{aligned} -(pu')' + qu - \lambda_n u = f \quad &\text{sur } [0, 1], \\ u(0) = u(1) = 0 & \end{aligned} \qquad (4.25)$$

est possible si et seulement si

$$\int_0^1 f(x)\, y_n(x)\, dx = 0,$$

et il existe une solution unique $z(x)$ telle que

$$\int_0^1 z(x)\, y_n(x)\, dx = 0 \qquad (4.26)$$

qui est de classe C^{k+2} $[0, 1]$.

Démonstration. Il résulte de [92] que le problème (4.25) admet une solution unique z continue sur $[0, 1]$ et vérifiant la condition (4.26), toutes les autres solutions du problème étant de la forme

$$u = z + \alpha y_n.$$

Le degré de régularité supérieur de la solution z est démontré par un procédé de récurrence usuel. En effet, on réduit le premier membre de (4.25) au terme

$$-(pz')' = (\lambda_n - q) z + f. \qquad (4.27)$$

Comme z, q et f sont continus, il en est de même du premier membre de (4.27). Par le théorème 1.1, on a donc $z \in C^2$ $[0, 1]$. Avec l'hypo-

thèse de p dans C^3 [0, 1], de $q \in C^2$ [0, 1] et de $f \in C^2$ [0, 1], le même théorème entraîne $z \in C^4$ [0, 1]. On poursuit le processus et on aboutit au résultat de régularité voulu.

En vertu de ce lemme, le problème (4.22), (4.23) est possible si et seulement si le second membre est orthogonal à la fonction propre y_n; cette condition est garantie pour σ_1 défini par la formule

$$\sigma_1 = \int_0^1 \left(\frac{(pv_0')'''}{24} + \frac{(pv_0''')'}{24} \right) y_n \, dx$$

qui utilise l'identité $v_0 = y_n$. Il est clair que σ_1 est indépendant de h et borné pour $v_0 \in C^4$ [0, 1].

Le problème (4.22), (4.23) ayant une infinité de solutions, on se pose maintenant le problème de choisir v_1. Toutes les solutions sont décrites par la formule

$$z_1 + \alpha_1 y_n,$$

z_1 étant la solution orthogonale à y_n et α_1 un paramètre réel.

On définit univoquement la constante α_1 moyennant la condition de normalisation (4.12) pour une fonction propre discrète. On porte le développement (4.16) dans (4.12):

$$\sum_{x \in \omega_h} \left[\sum_{s=0}^{l} h^{2s} v_s(x) \cdot \sum_{j=0}^{l} h^{2j} v_j(x) + h^r y_n^h(x) \eta^h(x) + \right.$$
$$\left. + h^r \eta^h(x) \sum_{s=0}^{l} h^{2s} v_s(x) \right] h = 1. \qquad (4.28)$$

On transforme le premier membre à l'aide du résultat bien connu relatif à la formule de quadrature des trapèzes (voir [23]).

LEMME 4.2. *On a pour toute fonction $f \in C^k$ [0, 1], $k \geq 3$, le développement*

$$\frac{h}{2} f(0) + \sum_{x \in \omega_h} f(x) h + \frac{h}{2} f(1) =$$
$$= \int_0^1 f(x)\, dx + \sum_{j=1}^{m} h^{2j} \frac{(-1)^j B_j}{(2j)!} \int_0^1 f^{(2j)}(x)\, dx + h^k g^h,$$

où $m = [(k-1)/2]$, B_j sont les nombres de Bernoulli (voir [49]):

$$B_0 = 1, \quad B_1 = \frac{1}{6}, \quad B_2 = \frac{1}{30}, \quad B_3 = \frac{1}{42}, \quad B_4 = \frac{1}{30}, \ldots,$$

et la constante g^h est uniformément bornée lorsque $h \to 0$.

On récrit (4.28) à la lumière du lemme 4.2:

$$\sum_{j=0} h^{2j}\left[\sum_{s=0}^{j}\sum_{k=0}^{l-j} h^{2k}\frac{(-1)^k B_k}{(2k)!}\int_0^1 (v_s(x)\, v_{j-s}(x))^{(2k)}\, dx + h^{r-2j}\, \nu_j^h\right] +$$

$$+\sum_{x\in\omega_h}\left[\sum_{j=l+1}^{2l} h^{2j}\sum_{k=j-l}^{l} v_k(x)\, v_{j-k}(x) + h^r\, y_n^h(x)\, \eta^h(x) + \right.$$

$$\left. + h^r\, \eta^h(x)\sum_{s=0} h^{2s}\, v_s(x)\right] h = 1, \qquad (4.29)$$

où

$$|\nu_j^h| \leqslant c_{12}, \qquad j = 0,\dots, l.$$

On identifie les coefficients des mêmes puissances de h dans (4.29). S'agissant de h^0, l'égalité

$$\int_0^1 v_0^2(x)\, dx = 1$$

est satisfaite automatiquement car $v_0 = y_n$. L'identification des coefficients de h^2 dans (4.29) conduit à la relation

$$2\int_0^1 v_0(x)\, v_1(x)\, dx - \frac{1}{12}\int_0^1 (v_0^2(x))''\, dx = 0.$$

Si l'on prend v_1 comme somme $z_1 + \alpha_1 y_n$, cette égalité et la condition d'orthogonalité de z_1 et y_n permettent de trouver le coefficient α_1:

$$\alpha_1 = \frac{1}{24}\int_0^1 (v_0^2(x))''\, dx. \qquad (4.30)$$

On pose

$$v_1 = z_1 + \alpha_1\, y_n. \qquad (4.31)$$

où z_1 est la solution orthogonale à y_n de (4.22), (4.23), et α_1 est défini par (4.30).

On obtient en identifiant les coefficients de h^4 dans (4.20)

$$-\sum_{0<k+s<2}\frac{\left(pv_{2-k-s}^{(2k+1)}\right)^{(2s+1)}}{(2k+1)!\,(2s+1)!\,4^{k+s}} + qv_2 - \sigma_0 v_2 + \sigma_1 v_1 + \sigma_2 v_0 \quad \text{sur } \omega_h.$$

Cette égalité donne lieu à l'équation pour v_2:

$$-(pv_2')' + q\, v_2 - \sigma_0\, v_2 =$$

$$= \sum_{1<k+s<2}\frac{\left(pv_{2-k-s}^{(2k+1)}\right)^{(2s+1)}}{(2k+1)!\,(2s+1)!\,4^{k+s}} + \sigma_1 v_1 + \sigma_2 v_0 \quad \text{sur} \quad [0,\ 1] \quad (4.32)$$

à laquelle nous adjoignons les conditions aux limites

$$v_2(0) = v_2(1) = 0 \tag{4.33}$$

(cf. (4.24)). Le problème (4.32), (4.33) admet une solution si le second membre de (4.32) est orthogonal à y_n. L'orthogonalité a lieu sous la condition nécessaire

$$\sum_{1 \leqslant k+s \leqslant 2} \int_0^1 \frac{\left(p(x)\, v^{(2k+1)}_{2-k-s}(x)\right)^{(2s+1)} y_n(x)}{(2k+1)!\,(2s+1)!\,4^{k+s}}\, dx + (\sigma_1 \alpha_1 + \sigma_2) \int_0^1 y_n^2(x)\, dx = 0,$$

d'où

$$\sigma_2 = -\sigma_1 \alpha_1 - \sum_{1 \leqslant k+s \leqslant 2} \int_0^1 \frac{\left(p(x)\, v^{(2k+1)}_{2-k-s}(x)\right)^{(2s+1)} y_n(x)}{(2k+1)!\,(2s+1)!\,4^{k+s}}\, dx. \tag{4.34}$$

Toutes les solutions du problème sont représentées par la formule

$$v_2 = z_2 + \alpha_2 y_n, \tag{4.35}$$

avec z_2 solution orthogonale à y_n. On choisit le coefficient α_2 en utilisant la condition d'identification des coefficients de h^4 dans (4.29):

$$\int_0^1 \left(2\, v_0(x)\, v_2(x) + v_1^2(x)\right) dx - \frac{1}{6} \int_0^1 \left(v_1(x)\, v_0(x)\right)'' dx + \\ + \frac{1}{720} \int_0^1 \left(v_0^2(x)\right)^{(4)} dx = 0.$$

Etant donnée la représentation (4.35) de la fonction v_2, on définit α_2 par la relation

$$\alpha_2 = -\frac{1}{2} \int_0^1 v_1^2(x)\, dx + \frac{1}{12} \int_0^1 \left(v_1(x)\, v_0(x)\right)'' dx - \frac{1}{1440} \int_0^1 \left(v_0^2(x)\right)^{(4)} dx.$$

On peut évidemment poursuivre.

Jusqu'à présent, nous avons agi dans l'hypothèse où les développements (4.16) et (4.17) existent à priori, et nous avons basé dessus un procédé de recherche de leurs termes.

On démontre que ces développements existent effectivement.

THÉORÈME 4.3. *Hypothèses* (4.3) *du problème de Sturm-Liouville* (4.1), (4.2), (4.4), (4.5). *Soit n un entier fixe. Il existe* $h_0 > 0$ *tel que la n-ième valeur propre* λ_n^h *du problème de Sturm-Liouville*

aux différences (4.9), (4.10), (4.12), (4.13) *et la fonction propre discrète associée* $y_n^h(x)$, $h \leq h_0$, *admettent les développements*

$$y_n^h(x) = y_n(x) + \sum_{s=1}^{l} h^{2s} v_s(x) + h^r \eta^h(x), \qquad x \in \bar{\omega}_h, \quad (4.36)$$

$$\lambda_n^h = \lambda_n + \sum_{s=1}^{l} h^{2s} \sigma_s + h^r \rho^h, \quad (4.37)$$

avec λ_n *la n-ième valeur propre du problème différentiel et* y_n *la fonction propre correspondante. Ici* $l = [(r-1)/2]$, *les fonctions* v_s *et les constantes* σ_s *sont indépendantes de* h, $v_s \in C^{r-2s+2}[0, 1]$ *et les restes sont évalués par*

$$|\rho^h| \leq c_9, \quad (4.38)$$

$$\max_{x \in \bar{\omega}_h} |\eta^h(x)| \leq c_{10}, \quad (4.39)$$

les constantes ne dépendant pas de h.

DÉMONSTRATION. On pose $v_0 = y_n$, $\sigma_0 = \lambda_n$ et on construit le système de problèmes différentiels

$$-(pv_j')' + qv_j - \sigma_0 v_j = f_j + \sum_{k=1}^{j-1} \sigma_k v_{j-k} + \sigma_j v_0 \quad (4.40)$$

sur $[0, 1]$,

$$v_j(0) = v_j(1) = 0, \quad (4.41)$$

où

$$f_j = \sum_{1 \leq k+s \leq j} \frac{\left(pv_{j-k-s}^{(2s+1)}\right)^{(2k+1)}}{4^{s+k}(2s+1)!\,(2k+1)!}, \quad (4.42)$$

$$\sigma_j = -\int_0^1 f_j(x)\, y_n(x)\, dx - \sum_{k=1}^{j-1} \sigma_k \int_0^1 v_{j-k}(x)\, y_n(x)\, dx, \quad (4.43)$$

$$j = 1, 2, \ldots, l.$$

On cherche une famille de fonctions v_j vérifiant la condition supplémentaire

$$\int_0^1 v_j(x)\, y_n(x)\, dx = -\sum_{\substack{1 \leq k+s \leq j \\ s \neq j}} \frac{(-1)^k B_k}{2(2k)!} \int_0^1 \left(v_s(x)\, v_{j-k-s}(x)\right)^{(2k)} dx, \quad (4.44)$$

$$j = 1, 2, \ldots, l.$$

Nous avons décrit plus haut un procédé de recherche de v_1 et v_2. Faisons l'hypothèse que les fonctions $v_0, \ldots, v_{j-1}$ satisfaisant aux équations (4.40) à (4.44) de numéros 1, 2, ..., $j-1$ sont définies, et $v_k \in C^{r+2-2k}\,[0, 1]$, $k = 0, 1, \ldots, j-1$. On cherche v_j pour les j-ièmes équations. On note que le second membre de (4.40) est par hypothèse une combinaison des fonctions connues $v_0, \ldots, v_{j-1}$ et appartient à la classe $C^{r-2j}\,[0, 1]$. Il est de plus orthogonal à y_n car on a, en vertu de la condition (4.43) et de la normalisation à l'unité de y_n et v_0,

$$\int_0^1 \Big(f_j(x) + \sum_{k=1}^{j-1} \sigma_k v_{j-k}(x) + \sigma_j v_0(x)\Big) y_n(x)\,dx = 0.$$

Aussi le j-ième problème (4.40) à (4.43) possède par le lemme 4.1 une solution dans l'espace fonctionnel $C^{r-2j+2}\,[0, 1]$, qui est unique si l'on lui impose d'être orthogonale à y_n. Soit z_j cette solution. On cherche v_j parmi les solutions $z_j + \alpha_j y_n$, le paramètre α_j étant réel défini par la formule

$$\alpha_j = - \sum_{\substack{1 \leqslant k+s \leqslant j \\ s \neq j}} \frac{(-1)^k B_k}{2\,(2k)!} \int_0^1 \big(v_s(x)\, v_{j-k-s}(x)\big)^{(2k)}\,dx \tag{4.45}$$

à partir de la condition supplémentaire (4.44). La continuité des fonctions v_k, $k = 0, 1, \ldots, j-1$, fait conclure au caractère fini de α_j. La fonction

$$v_j = z_j + \alpha_j y_n$$

est la fonction cherchée et est dans $C^{r-2j+2}\,[0, 1]$.

On procède de même pour toutes les $v_0, v_1, \ldots, v_l$.

Soit λ_n^h la n-ième valeur propre discrète du problème (4.9), (4.10) et y_n^h la fonction propre associée remplissant les conditions de normalisation (4.12), (4.13). Il est évident que h est inférieur à $1/n$. On suppose connues les fonctions y_n^h, v_j et les constantes λ_n^h, σ_j on définit η^h et ρ^h par les formules

$$\eta^h = h^{-r}\Big(y_n^h - \sum_{j=1}^{l} h^{2j} v_j\Big) \quad \text{sur} \quad \bar{\omega}_h. \tag{4.46}$$

$$\rho^h = h^{-r}\Big(\lambda_n^h - \sum_{j=1}^{l} h^{2j} \sigma_j\Big) \tag{4.47}$$

et on montre leur propriété d'être bornés pour $h \to 0$.

On exprime y_n^h et λ_n^h à partir de (4.46), (4.47) et on les porte dans l'équation (4.9). On est évidemment conduit à (4.20). Avec l'équation (4.1) juste pour v_0 et σ_0, on supprime dans l'identité (4.20) les termes en h^0. Le procédé de définition des fonctions v_j (les équations (4.40)) et des constantes σ_j (l'égalité (4.43)) entraîne la réduction de tous les termes en $h^2, h^4, \dots, h^{2l}$. On divise les termes restants par h^r et on forme avec les groupes suivants:

$$-(p\eta_{\bar{x}}^h)_x + q\eta^h - \lambda_n^h\eta^h =$$

$$= \sum_{j=0}^{l} \mu_j^h + h^{2l+2-r} \sum_{j=0}^{l-1} h^{2j} \sum_{k=j+1} \sigma_k v_{j-k} + \rho^h \sum_{j=0} h^{2j} v_j. \qquad (4.48)$$

On note que le second membre comprend les expressions indépendantes de η^h. Cette identité étant juste pour η^h réelle est résoluble par rapport à η^h. La condition de possibilité est donnée par le

Lemme 4.4. *Soit λ_n^h une valeur propre discrète et y_n^h la fonction propre correspondante. Le problème aux différences*

$$-(pw_{\bar{x}}^h)_x + qw^h - \lambda_n^h w^h = f \quad \text{sur} \quad \omega_h, \qquad (4.49)$$

$$w^h(0) = w^h(1) = 0$$

possède une solution si et seulement si

$$\sum_{x\in\omega_h} f(x)\, y_n^h(x)\, h = 0, \qquad (4.50)$$

et l'ensemble des solutions s'écrit sous forme de somme

$$w^h = z^h + \alpha y_n^h \quad \text{sur} \quad \bar{\omega}_h, \qquad (4.51)$$

où α est un paramètre réel indépendant et z^h la solution du problème (4.49), (4.50) *de norme minimum*

$$\sum_{x\in\omega_h} (z^h(x))^2\, h.$$

L'estimation

$$\max_{x\in\omega_h} |f(x)| \leqslant c_{12} \qquad (4.52)$$

entraîne de plus

$$\max_{x\in\bar{\omega}_h} |z^h(x)| \leqslant c_{13}. \qquad (4.53)$$

Démonstration. On élimine dans (4.49) les valeurs nulles $w^h(0)$, $w^h(1)$ et on passe aux notations matricielles, il vient

$$(A - \lambda_n^h I)\, W^h = F, \qquad (4.54)$$

avec I la matrice unité et A la matrice symétrique de (4.11). Les vecteurs W^k et F sont définis par

$$W^k = \begin{bmatrix} w^k(x_1) \\ \cdot \\ \cdot \\ \cdot \\ w^k(x_{N-1}) \end{bmatrix}, \qquad F = \begin{bmatrix} f(x_1) \\ \cdot \\ \cdot \\ \cdot \\ f(x_{N-1}) \end{bmatrix}.$$

Comme λ_n^k est valeur propre de A, on a

$$\det (A - \lambda_n^k I) = 0.$$

Le système (4.54) est possible si et seulement si le second membre F est orthogonal au noyau de la matrice $(A - \lambda_n^k I)$ dans l'espace euclidien de dimension $N - 1$. Le noyau est de dimension un et ses éléments sont des vecteurs αY_n^k, α étant un paramètre réel quelconque et Y_n^k un vecteur propre de A associé à la valeur propre λ_n^k. Ainsi, le problème (4.49) est possible sous la condition nécessaire et suffisante $F^T Y_n^k = 0$. On multiplie cette relation par h et on est conduit à (4.50). Ainsi, il y a existence, et toute solution de (4.54) est de la forme

$$W^k = Z^k + \alpha Y_n^k,$$

où Z^k est la solution normale (voir [136]) du système, i.e. la solution de norme minimum $(Z^k)^T Z^k$. L'expression

$$\sum_{x \in \omega_h} (z^k(x))^2 h$$

coïncide avec cette quantité au facteur h près si la fonction discrète $z^k(x)$ est définie par la formule

$$Z^k = \begin{bmatrix} z^k(x_1) \\ \cdot \\ \cdot \\ \cdot \\ z^k(x_{N-1}) \end{bmatrix}.$$

Ainsi, on a démontré la représentation (4.51) des solutions. On prouve la majoration (4.53) en mettant le second membre de l'équation (4.49) sous la forme

$$f(x) = \sum_{\substack{i=1 \\ i \neq n}}^{N-1} \beta_i y_i^k (x),$$

ce qui est possible en vertu de la condition (4.50). Cela étant, les paramètres réels β_i sont évalués par

$$\sum_{\substack{i=1 \\ i\neq n}}^{N-1} \beta_i^2 = \sum_{x\in\omega_h} f^2(x)\, h \leqslant c_{12}^2. \tag{4.55}$$

Dans ce cas, la solution $z^h(x)$ s'écrit

$$z^h(x) = \sum_{\substack{i=1 \\ i\neq n}}^{N-1} \frac{\beta_i}{\lambda_i^h - \lambda_n^h}\, y_i^h(x).$$

On apprécie la différence des valeurs propres à partir du degré de proximité des valeurs propres du problème discrétisé et de celles du problème différentiel. On se sert à cet effet du résultat suivant sur la convergence des valeurs et fonctions propres discrètes vers leurs homologues du problème différentiel.

LEMME 4.5 (voir [41]). *On suppose les coefficients p et q du problème* (4.1), (4.2) *appartenir à C^2* [0, 1]. *La n-ième solution du problème aux différences* (4.9), (4.10) *converge vers la solution du problème différentiel et la convergence est d'ordre* 2:

$$\max_{x\in[0,1]} | y_n(x) - y_n^h(x) | \leqslant c_{14} h^2,$$

$$| \lambda_n - \lambda_n^h | \leqslant c_{15} h^2,$$

où les constantes c_{14}, c_{15} sont indépendantes de h (mais dépendent de n).

Considérons les valeurs propres du problème différentiel (4.1), (4.2). Elles sont rangées dans l'ordre de croissance, si bien que la valeur propre le plus proche de λ_n est, ou bien λ_{n+1}, ou bien λ_{n-1}. On désigne par a l'écart minimal:

$$a = \min \{ | \lambda_n - \lambda_{n-1} |, | \lambda_n - \lambda_{n+1} | \}$$

et on note que

$$| \lambda_n - \lambda_i | \geqslant a > 0 \qquad \forall_i \neq n.$$

Passons aux valeurs propres du problème aux différences. La valeur propre le plus proche de λ_n^h est, ou bien λ_{n+1}^h, ou bien λ_{n-1}^h. Supposons que ce soit λ_{n+1}^h, ce qui ne restreint en rien la généralité de l'exposé. On a

$$| \lambda_n^h - \lambda_i^h | \geqslant | \lambda_n^h - \lambda_{n+1}^h | \qquad \forall\, i \neq n. \tag{4.56}$$

D'après le lemme 4.5, λ_n^h et λ_{n+1}^h convergent vers λ_n et λ_{n+1} respectivement avec h tendant vers 0. On trouve donc $\varepsilon_1 > 0$ tel que

$$|\lambda_n - \lambda_n^h| \leqslant \frac{a}{4}, \qquad |\lambda_{n+1} - \lambda_{n+1}^h| \leqslant \frac{a}{4}$$

$\forall\ h < \varepsilon_1$. Ainsi,

$$|\lambda_n^h - \lambda_{n+1}^h| = |\lambda_n^h - \lambda_n + \lambda_n - \lambda_{n+1} + \lambda_{n+1} - \lambda_{n+1}^h| \geqslant$$
$$\geqslant |\lambda_n - \lambda_{n+1}| - |\lambda_n^h - \lambda_n| - |\lambda_{n+1} - \lambda_{n+1}^h| \geqslant \frac{a}{2}.$$

Vu (4.56), on aboutit à l'inégalité

$$|\lambda_n^h - \lambda_i^h| \geqslant \frac{a}{2} \qquad \forall\ i \neq n. \qquad \forall\ h \leqslant \varepsilon_1.$$

Cette relation et l'estimation (4.55) conduisent à

$$\sum_{x \in \omega_h} (z^h(x))^2 h \leqslant \sum_{\substack{i=1 \\ i \neq n}}^{N-1} \frac{4\beta_i^2}{a^2} \leqslant \frac{4c_{12}^2}{a^2}.$$

On observe en outre que z^h est solution du problème

$$\begin{aligned} &-(pz_{\bar{x}}^h)_x + qz^h = f + \lambda_n^h z^h \quad \text{sur} \quad \omega_h, \\ &z^h(0) = z^h(1) = 0; \end{aligned} \tag{4.57}$$

avec $g = f + \lambda_n^h z^h$ borné. On a notamment, par suite de (4.14),

$$\Big(\sum_{x \in \omega_h} g^2(x)\, h\Big)^{1/2} \leqslant \Big(\sum_{x \in \omega_h} f^2(x)\, h\Big)^{1/2} + |\lambda_n^h| \Big(\sum_{x \in \omega_h} (z^h(x))^2\Big)^{1/2} \leqslant$$
$$\leqslant c_{12} + \frac{2}{a} c_{12}\, c_7\, n^2.$$

La solution z^h est évaluée par

$$\max_{x \in [0,1]} |z^h(x)| \leqslant \frac{1}{c_1} \Big(\sum_{x \in \omega_h} g^2(x)\, h\Big)^{1/2}$$

de [43]). Si l'on pose

$$c_{13} = \frac{c_{12}}{c_1}\left(1 + \frac{2}{a} c_7 n^2\right),$$

l'estimation (4.53) découle de deux dernières relations, et le lemme 4.4 se trouve démontré.

On reprend l'équation (4.48).

On vérifie que la fonction discrète η^h s'annule aux extrémités du segment [0, 1]. En effet,

$$\eta^h(b) = h^{-r}\Big(y_n^h(b) - \sum_{j=0} h^{2j} v_j(b)\Big) = 0 \tag{4.58}$$

si $b = 0$ ou $b = 1$ (en vertu des conditions aux limites homogènes (4.2), (4.10) et (4.41)). Aux termes du lemme 4.4, le second membre de (4.48) doit être orthogonal (au sens discrétisé) à y_n^h, si bien que

$$\rho^h = -\frac{\displaystyle\sum_{x\in\omega_h}\Big(\sum_{j=0}^{l}\mu_j^h(x) + h^{2l+2-r}\sum_{j=0}^{l-1}h^{2j}\sum_{k=j+1}^{l}\sigma_k\, v_{j-k}(x)\Big)\, y_n^h(x)h}{\displaystyle\sum_{x\in\omega_h} y_n^h(x)\sum_{j=0}^{l}h^{2j}\, v_j(x)\, h}\,.$$

Le numérateur est évalué en module moyennant l'inégalité de Cauchy-Bouniakovski :

$$\Big(\sum_{x\in\omega_h}\Big(\sum_{j=0}\mu_j^h(x) + h^{2l+2-r}\sum_{j=0}^{l-1}h^{2j}\sum_{k=j+1}^{l-1}\sigma_k\, v_{j-k}(x)\Big)^2 h\Big)^{1/2} \leqslant$$
$$\leqslant c_{11}(l+1) + c_{16}\, c_{17}\, l^2,$$

avec la constante c_{11} de (4.19),

$$c_{16} = \max_{0\leqslant k\leqslant l}|\sigma_k|, \qquad c_{17} = \max_{\substack{x\in[0,1]\\ 0\leqslant j\leqslant l}}|v_j(x)|. \tag{4.59}$$

On demande d'évaluer le dénominateur. Il résulte du lemme 4.5 que y_n^h converge vers y_n pour $h\to 0$ uniformément sur $\bar\omega_h$. Aussi, on a, à partir d'un certain $\varepsilon_2 > 0$,

$$\max_{x\in\omega_h}|y_n^h(x) - y_n(x)| \leqslant 1/4 \qquad \forall\, h \leqslant \varepsilon_2.$$

Les fonctions v_j étant continues sur [0, 1] sont bornées. Donc

$$\lim_{h\to 0}\max_{x\in[0,1]}\Big|\sum_{j=1}^{l}h^{2j}\, v_j(x)\Big| = 0.$$

On prend $\varepsilon_3 > 0$ tel que

$$\max_{x\in[0,1]}\Big|\sum_{j=1}^{l}h^{2j}\, v_j(x)\Big| \leqslant 1/4 \qquad \forall\, h \leqslant \varepsilon_3.$$

D'où

$$\sum_{x\in\omega_h} y_n^h(x) \sum_{j=0}^{l} h^{2j} v_j(x)\, h =$$

$$= \sum_{x\in\omega_h} y_n^h(x) \left(y_n^h(x) - y_n^h(x) + v_0(x) + \sum_{j=1}^{l} h^{2j} v_i(x) \right) h =$$

$$= 1 - \sum_{x\in\omega_h} y_n^h(x) \left(y_n^h(x) - v_0(x) \right) h +$$

$$+ \sum_{x\in\omega_h} y_n^h(x) \sum_{j=1}^{l} h^{2j} v_j(x)\, h \geqslant 1 - \frac{1}{2} \sum_{x\in\omega_h} | y_n^h(x) |\, h \geqslant$$

$$\geqslant 1 - \frac{1}{2} \Big(\sum_{x\in\omega_h} (y_n^h(x))^2 h \Big)^{1/2} = \frac{1}{2}, \qquad h \leqslant \min\{\varepsilon_2, \varepsilon_3\}.$$

Aussi on a pour $h \leqslant \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$

$$| \rho^h | \leqslant c_{18} = 2(l+1)\, c_{11} + 2\, l^2 c_{16} c_{17}. \tag{4.60}$$

Avec ρ^h ainsi choisi, le second membre de (4.48) est uniformément borné par la constante

$$\frac{c_{18}}{2} + c_{18} (l + 1)\, c_{17},$$

si bien que par le lemme 4.4 la solution orthogonale à y_n^h du problème (4.48), (4.58) est également bornée:

$$\max_{x\in\omega_h} | z^h(x) | \leqslant c_{19}, \; h \leqslant \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}. \tag{4.61}$$

Comme la solution générale du problème s'écrit

$$z^h(x) + \alpha^h y_n^h(x),$$

il est donc nécessaire de trouver la constante α^h et d'établir sa propriété d'être bornée. On obtient α^h moyennant la condition de normalisation (4.12). Pour cela, on définit y_n^h à partir de (4.46) et on substitue dans (4.12). On est évidemment conduit à l'identité (4.29). On utilise la condition de normalisation (4.4) vraie pour v_0, on ré-

duit les termes en h^0 de (4.29) et on partage les termes restants en plusieurs groupes :

$$\sum_{j=1}^{l} h^{2j}\left(\sum_{0<k+s<j} \frac{(-1)^k B_k}{2(2k)!}\int_0^1 (v_s(x)\, v_{j-k-s}(x))^{2k} dx\right) +$$

$$+ h^r \sum_{j=0}^{l} \nu_j^h + \sum_{x\in\omega_h}\left(\sum_{j=l+1}^{2l} h^{2j} \sum_{k=j-l} v_k(x)\, v_{j-k}(x)\right) h +$$

$$+ h^r \sum_{x\in\omega_h} \eta^h(x)\left(y_n^h(x) + \sum_{s=0}^{l} h^{2s} v_s(x)\right) h = 0.$$

On voit que la condition (4.44) détermine la réduction de tous les termes en $h^2, h^4, \ldots, h^{2l}$. On divise les termes restants par h^r, il vient la relation

$$\sum_{x\in\omega_h} \eta^h(x)\left(y_n^h(x) + \sum_{s=0}^{l} h^{2s} v_s(x)\right) h = \varkappa^h, \tag{4.62}$$

où

$$\varkappa^h = -\sum_{j=0}^{l} \nu_j^h - h^{2l+2-r} \sum_{x\in\omega_h} \sum_{j=0}^{l-1} h^{2j} \sum_{k=j+1}^{l} v_k(x)\, v_{j-k}(x)\, h$$

est une quantité bornée parce que

$$|\varkappa^h| \leqslant c_{20} = (l+1)\, c_{12} + c_{17}^2\, l^2. \tag{4.63}$$

La constante c_{12} est définie à partir de la condition (4.29) et la constante c_{17} est celle de (4.59).

La formule (4.46) donnant η^h permet de ramener l'égalité (4.62) à la forme

$$2\sum_{x\in\omega_h} \eta^h(x)\, y_n^h(x)\, h - h^r \sum_{x\in\omega_h} (\eta^h(x))^2 h - \varkappa^h = 0.$$

On fait la substitution $\eta^h = z^h + \alpha^h y_n^h$, il vient l'équation (par rapport à α^h)

$$2\alpha^h - h^r \sum_{x\in\omega_h} (z^h(x))^2 h - (\alpha^h)^2 h^r - \varkappa^h = 0.$$

On note qu'elle admet deux racines

$$\alpha_1^h = \frac{1 + \sqrt{1 - h^r\left(\varkappa^h + h^r \sum_{x \in \omega_h} (z^h(x))^2 h\right)}}{h^r},$$

$$\alpha_2^h = \frac{1 - \sqrt{1 - h^r\left(\varkappa^h + h^r \sum_{x \in \omega_h} (z^h(x))^2 h\right)}}{h^r}.$$

La fonction η^h est définie de façon unique, si bien qu'on prend pour paramètre α^h une seule des racines. Comme $z^h(x)$ est uniformément bornée (voir (4.61)) par la constante c_{19} et $\varkappa^h$ par c_{20} (voir (4.63),

$$\left| \varkappa^h + h^r \sum_{x \in \omega_h} (z^h(x))^2 h \right| \leqslant c_{20} + c_{19}^2,$$

et on a, à partir de

$$\varepsilon_4 = \sqrt[r]{\frac{3}{4(c_{20} + c_{19}^2)}},$$

l'estimation

$$\alpha_1^h \geqslant \frac{1 + \sqrt{1/4}}{h^r} = \frac{3}{2h^r} \qquad \forall h \leqslant \varepsilon_4. \tag{4.64}$$

On pose $\alpha^h = \alpha_1^h$. On a par définition de η^h

$$y_n^h(x_1) = y_n(x_1) + \sum_{s=1}^{l} h^{2s} v_s(x_1) + h^r \eta^h(x_1),$$

et la relation

$$\eta^h = z^h + \alpha_1^h y_n^h$$

fournit

$$(1 - h^r \alpha_1^h)\, y_n^h(x_1) = y_n(x_1) + \sum_{s=1}^{l} h^{2s} v_s(x_1) + h^r z^h(x_1). \tag{4.65}$$

Du moment que $v_s(x_1)$ et $z^h(x_1)$ sont bornés, on aboutit à une contradiction avec les conditions de normalisation (4.5) et (4.13). En effet,

$$y_n(x_1) = y_n(0) + h y_n'(0) + \frac{h^2}{2} y_n''(\xi),$$

où $\xi \in (0, x_1)$. Puisque $y'_n(0) > 0$ et $y_n(0) = 0$, on garantit à partir d'un certain ε_5 que $y_n(x_1) > 0$ quand $h \to 0$. Aussi $y_n(x_1)$ devient, à partir d'un certain ε_6, $h \leqslant \varepsilon_6$, le terme prépondérant du second membre de (4.65) et en détermine le signe. Quant au premier membre, il est un nombre négatif du moment que l'estimation (4.64) entraîne $(1 - h^r \alpha_1^h) \leqslant -1/2$ et que $y_n^h(x_1) > 0$ par la condition de normalisation.

Ainsi, le fait d'avoir posé $\alpha^h = \alpha_1^h$ nous a conduit à des signes opposés. Aussi on pose $\alpha^h = \alpha_2^h$. On vérifie la propriété de borne. On a

$$|\alpha_2^h| = \begin{cases} \dfrac{\sqrt{1+\beta} - 1}{h^r} & \text{si} \quad \beta \geqslant 0, \\ \dfrac{1 - \sqrt{1+\beta}}{h^r} & \text{si} \quad 0 > \beta > -1, \end{cases}$$

où

$$\beta = -h^r\Big(\varkappa^h + h^r \sum_{x \in \omega_h} (z^h(x))^2 h\Big).$$

On se place dans le cas $\beta \geqslant 0$. L'inégalité

$$\sqrt{1+\beta} \leqslant 1 + \beta/2$$

entraîne

$$|\alpha_2^h| \leqslant -\frac{1}{2}\varkappa^h - \frac{h^r}{2} \sum_{x \in \omega_h} (z^h(x))^2 h \leqslant \frac{c_{20} + c_{19}^2}{2}.$$

Si $\beta < 0$, alors

$$\beta \geqslant -\frac{3}{4} \qquad \forall\, h \leqslant \varepsilon_4 = \sqrt[r]{\frac{3}{4(c_{20} + c_{19}^2)}}$$

et

$$\sqrt{1+\beta} \geqslant 1 + \frac{2}{3}\beta.$$

Donc

$$|\alpha_2^h| \leqslant -\frac{2}{3}\frac{\beta}{h^r} = \frac{2}{3}\Big(\varkappa^h + h^r \sum_{x \in \omega_h} (z^h(x))^2 h\Big) \leqslant \frac{2}{3}(c_{20} + c_{19}^2).$$

Ainsi, on a dans les deux cas l'estimation

$$|\alpha_2^h| \leqslant \frac{2}{3}(c_{20} + c_{19}^2).$$

Compte tenu de (4.15), on aboutit à l'inégalité cherchée

$$|\gamma_i^h(x)| \leqslant c_8\, n^{1/2} \frac{2}{3}(c_{20} + c_{19}^2) + c_{19} \qquad \forall h \leqslant h_0,$$

où

$$h_0 = \min_{1 \leqslant i \leqslant 5} \varepsilon_i,$$

ce qui achève la démonstration du théorème 4.3.

REMARQUE. Les constantes c_9 et c_{10} des estimations (4.38), (4.39) dépendent essentiellement du numéro de la valeur et fonction propre (elles augmentent avec le numéro). Cette dépendance par rapport à n peut être mise en évidence tout au long de la démonstration du théorème, mais l'exemple numérique à la fin du paragraphe l'illustre mieux.

On note qu'avec les développements (4.36), (4.37), on peut réaliser l'extrapolation de Richardson, et la combinaison linéaire est dans les deux cas affectée de mêmes poids. On se propose de décrire l'algorithme d'extrapolation pour la recherche de la n-ième valeur propre et de la n-ième fonction propre correspondante.

Supposons qu'on est dans les conditions de régularité (4.3) et posons $s = [(r-1)/2]$. On construit pour $s+1$ entiers $N_k = kL$ les réseaux $\bar{\omega}_{h_k}$ de pas $h_k = 1/N_k$, $k = 1, \ldots, s+1$, et les problèmes de Sturm-Liouville discrets (4.9) à (4.11). On cherche dans chaque problème la n-ième valeur propre $\lambda_n^{h_k}$ et la fonction propre associée, et on norme chaque fonction de façon à remplir les conditions (4.12), (4.13). Les $s+1$ fonctions propres discrètes $y_n^{h_k}$ sont définies sur le réseau $\bar{\omega}_{h_1}$. On forme la combinaison linéaire

$$Y_n = \sum_{k=1}^{s+1} \gamma_k\, y_n^{h_k} \quad \text{sur} \quad \bar{\omega}_{h_1}, \tag{4.66}$$

où les poids

$$\gamma_k = \frac{2(-1)^{s-k+1}\, k^{2s+2}}{(s-k+1)!\,(s+k+1)!}. \tag{4.67}$$

On a de même pour les valeurs propres discrètes et les mêmes poids:

$$\Lambda_n = \sum_{k=1}^{s+1} \gamma_k\, \lambda_n^{h_k}. \tag{4.68}$$

THÉORÈME 4.6. *On suppose qu'on est dans les hypothèses du théorème 4.3 pour un entier n fixé. La valeur propre discrète corrigée*

(4.68) *et la fonction propre discrète corrigée* (4.66) *admettent les estimations*

$$\max_{x\in\bar{\omega}_{h_1}} |Y_n(x) - y_n(x)| \leqslant h_1^r c_{21}, \tag{4.69}$$

$$|\Lambda_n - \lambda_n| \leqslant h_1^r c_{22}, \tag{4.70}$$

où les constantes c_{21}, c_{22} *ne dépendent pas de* h.

DÉMONSTRATION. Selon le théorème 4.3, les valeurs propres admettent les développements

$$\lambda_n^{h_k} = \lambda_n + \sum_{j=1}^{s} h_k^{2j}\sigma_j + h_k^r \rho^{h_k}, \qquad k = 1, 2, \ldots, s.$$

On fait la somme avec les poids γ_k, il vient l'égalité

$$\Lambda_n = \sum_{k=1}^{s+1}\gamma_k \lambda_n + \sum_{j=1}^{s}\left(\sum_{k=1}^{s+1}\gamma_k h_k^{2j}\right)\sigma_j + \sum_{k=1}^{s+1}\gamma_k h_k^r \rho^{h_k}. \tag{4.71}$$

Aux termes du lemme 2.2, § 7.2, les poids γ_k vérifient

$$\sum_{k=1}^{s+1}\gamma = 1, \qquad \sum_{k=1}^{s+1}\gamma_k \frac{1}{k^{2j}} = 0, \qquad j = 1, \ldots, s,$$

ou, ce qui revient au même, les égalités

$$\sum_{k=1}^{s+1}\gamma_k h_k^{2j} = 0, \qquad j = 1, \ldots, s.$$

On en tient compte dans (4.71) qui devient

$$\Lambda_n = \lambda_n + \sum_{k=1}^{s+1}\gamma_k h_k^r \rho^{h_k},$$

d'où

$$|\Lambda_n - \lambda_n| \leqslant \sum_{k=1}^{s+1} |\gamma_k| \; |h_k^r| \, \rho^{h_k}|. \tag{4.72}$$

La quantité $|\rho^{h_k}|$ est évaluée par la constante c_9 de (4.38), et $|\gamma_k| h_k^r$ devient

$$|\gamma_k| h_k^r = h_1^r \frac{2k^{2s+2-r}}{(s-k+1)!(s+k+1)!}.$$

Aussi l'inégalité (4.72) s'écrit

$$| \Lambda_n - \lambda_n | \leqslant c_9 \, h_1^r \sum_{k=1}^{s+1} \frac{2 \, k^{2s+2-r}}{(s-k+1)! \, (s+k+1)!},$$

ce qui équivaut à (4.70), où

$$c_{22} = c_9 \sum_{k=1}^{s+1} \frac{2 \, k^{2s+2-r}}{(s-k+1)! \, (s+k+1)!}.$$

On traite de même le développement (4.36), ce qui fournit l'inégalité (4.69). Le théorème se trouve démontré.

Illustrons ces résultats par des exemples numériques. On pose $\alpha = (e^2 - 1)/2$ et on considère le problème de trouver les valeurs propres et les fonctions propres:

$$\begin{gathered} -\big((2 \, \alpha x + 1)^2 \, u' \big)' = \lambda u \quad \text{sur} \quad [0, 1], \\ u(0) = u(1) = 0. \end{gathered} \tag{4.73}$$

La solution du problème est la suite de valeurs propres

$$\lambda_n = (1 + \pi^2 n^2) \, \alpha^2, \qquad n = 1, 2, \ldots,$$

et le système orthonormé correspondant de fonctions propres

$$y_n(x) = \sqrt{\frac{2 \alpha}{2 \alpha x + 1}} \sin \left(\frac{n \pi}{2} \ln (2 \, \alpha x + 1) \right).$$

On fixe n et on construit en coordonnées logarithmiques les graphes des fonctions

$$\xi_\lambda(N) = | \lambda_n^h - \lambda_n |, \qquad \xi_y(N) = \max_{x \in \omega_h} | y_n^h(x) - y_n(x) |.$$

On trouve pour plusieurs N et les réseaux ω_h et $\omega_{h/2}$, $h = 1/N$, les valeurs extrapolées Λ_n et Y_n (on utilise les formules (4.66), (4.68) pour $s = 1$), et on établit les erreurs dont ces valeurs sont affectées:

$$\zeta_\lambda(3N) = | \Lambda_n - \lambda_n |, \qquad \zeta_y(3N) = \max_{x \in \bar{\omega}_h} | Y_n(x) - y_n(x) |.$$

Les erreurs dépendent de $3N$ car la construction de Λ_n et Y_n coûte autant que la résolution du problème approché (4.9), (4.10) sur un réseau de pas $1/(3N)$.

La figure 3.1 donne le graphe correspondant pour les premières quatre valeurs propres et la figure 3.2 celui pour les fonctions propres associées. Les graphes montrent nettement la dépendance de la précision des solutions approchées vis-à-vis de numéro de la valeur et fonction propre.

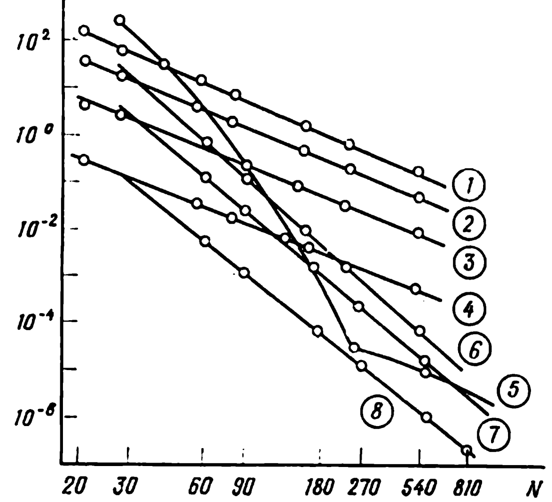


Fig. 3.1. Erreurs sur les valeurs propres discrètes et les valeurs propres extrapolées du problème (4.73)

Erreurs sur les valeurs propres discrètes du problème (4.9), (4.10): *1* — 4ième v.p.; *2* — 3ième v.p.; 3—2ième v.p; *4* — 1re v.p. Erreurs sur les valeurs propres extrapolées (4.68); *5* — 4ième v.p.; *6* — 3ième v.p.; *7* — 2ième v.p.; *8* — 1re v.p.

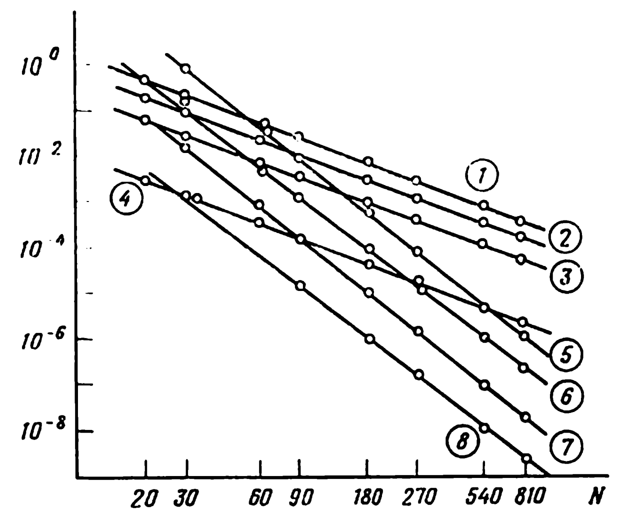


Fig. 3.2. Erreurs maxima sur les fonctions propres discrètes et les fonetions propres extrapolées du problème (4.73)

Erreurs sur le fonctions propres discrètes du problème (4.9), (4.10): *1* — 4ième f.p.; *2*—3ième f.p; *3* — 2ième f.p.; *4* — 1re f.p. Erreurs sur les fonctions propres extrapolées (4.66); *5* — 4ième f.p.; *6*— 3ième f.p.; *7* — 2ième f.p; *8* — 1re f.p.

3.5. Amélioration de la précision dans la méthode des éléments finis

Depuis plusieurs années, on voit paraître de nombreux ouvrages consacrés aux méthodes variationnelles des différences finies pour les problèmes de la physique mathématique. La méthode des éléments finis en est la plus répandue (voir par exemple [37], [81], [112], [132]). S'agissant des problèmes auto-adjoints, elle est basée, on le sait, sur la minimisation d'une fonctionnelle variationnelle et sur la représentation de la solution par une combinaison de fonctions d'essai (d'éléments finis) choisies de façon spéciale. Si l'on se borne aux fonctions « chapeaux pointus », on est conduit de règle à des équations simples à trois points pour déterminer les coefficients inconnus de la solution, et ces équations sont en général exactes à l'ordre 1 ou 2. Avec des éléments finis plus évolués (ou plus réguliers), on obtient des équations à plusieurs points qui sont plus compliquées mais aussi plus exactes. On note de plus que dans le cas d'éléments finis réguliers et de l'équation différentielle à coefficients discontinus, l'algorithme de construction des équations aux différences se complique singulièrement. Nous nous servirons donc, pour bâtir les équations variationnelles, aux éléments finis simples, auquel cas la combinaison linéaire de solutions associées à plusieurs réseaux successifs donne une solution de haute précision. On observe que ce procédé présente l'avantage sérieux de résoudre tous les problèmes élémentaires de manière uniforme (seul varie le paramètre de discrétisation). Cette question fait l'objet du présent paragraphe.

Voici un problème plus simple que (1.1):

$$-u'' + q(x)\,u = f(x), \qquad x \in (0,\ 1), \tag{5.1}$$

$$u(0) = u_0, \qquad u(1) = u_1 \tag{5.2}$$

à coefficients

$$f,\ q \in C^r[0,\ 1], \qquad q \geqslant 0 \quad \text{sur} \quad [0,\ 1], \tag{5.3}$$

où $r \geqslant 2$ est un entier. On trouve dans [78] plusieurs techniques permettant de ramener l'équation (1.1) à (5.1).

La résolution numérique du problème proposé consiste à associer à chaque nœud du réseau régulier ω_h une fonction d'essai définie sur [0, 1] par la formule

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h} & \text{si} \quad x - x_i \in (-h,\ 0], \\ \dfrac{x_{i+1} - x}{h} & \text{si} \quad x - x_i \in (0,\ h), \\ 0 & \text{dans les cas restants.} \end{cases} \tag{5.4}$$

On fixe i, $1 \leqslant i \leqslant N - 1$, on multiplie (5.1) par $\varphi_i(x)$ et on intègre sur $[0, 1]$, il vient

$$\int_0^1 (-u'' \varphi_i + qu \varphi_i)\, dx = \int_0^1 f \varphi_i \, dx.$$

On intègre par parties et on utilise la propriété $\varphi_i(x_i \pm h) = 0$:

$$\int_0^1 (u' \varphi_i' + qu \varphi_i)\, dx = \int_0^1 f \varphi_i \, dx. \tag{5.5}$$

Avec les notations simplificatrices

$$(v, w) = \int_0^1 v(x)\, w(x)\, dx,$$

$$[v, w] = \int_0^1 (v' w' + qvw)\, dx,$$

on écrit l'identité (5.5)

$$[u, \varphi_k] = (f, \varphi_k), \qquad k = 1, \ldots, N - 1. \tag{5.6}$$

L'idée de la méthode des éléments finis est de chercher la solution approchée $u^h(x)$ comme

$$u^h(x) = \sum_{i=0}^{N} \alpha_i \varphi_i(x). \tag{5.7}$$

Ici α_i sont un jeu de constantes définies à partir des égalités obtenues par substitution $u^h(x) = u(x)$ dans (5.6). On a

$$[u^h, \varphi_k] = (f, \varphi_k) \tag{5.8}$$

ou

$$\sum_{i=0}^{N} \alpha_i [\varphi_i, \varphi_k] = (f, \varphi_k), \qquad k = 1, \ldots, N - 1.$$

Les conditions aux limites sont vérifiées si l'on pose

$$\alpha_0 = u_0, \qquad \alpha_N = u_1. \tag{5.9}$$

On forme ainsi un système d'équations en α_i dont l'écriture matricielle est

$$A\alpha = \begin{bmatrix} b_1 & c_1 & & & 0 \\ a_2 & b_2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \\ 0 & & & \cdot \quad \cdot & e_{N-2} \\ & & & a_{N-1} & b_{N-1} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{N-2} \\ \alpha_{N-1} \end{bmatrix} = \begin{bmatrix} g_1 - a_1 u_0 \\ g_2 \\ \cdot \\ \cdot \\ \cdot \\ g_{N-2} \\ g_{N-1} - e_{N-1} u_1 \end{bmatrix}, \quad (5.10)$$

où

$$a_i = [\varphi_{i-1}, \varphi_i] = -\frac{1}{h} + \int_{x_{i-1}}^{x_i} \frac{x_i - x}{h} \frac{x - x_{i-1}}{h} q(x)\, dx,$$

$$e_i = a_{i+1} = [\varphi_i, \varphi_{i+1}],$$

$$b_i = [\varphi_i, \varphi_i] = \frac{2}{h} + \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1})^2}{h^2} q(x)\, dx + \int_{x_i}^{x_{i+1}} \frac{(x_{i+1} - x)^2}{h^2} q(x)\, dx,$$

$$g_i = (f, \varphi_i) = \int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{h} f(x)\, dx + \int_{x_i}^{x_{i+1}} \frac{x_{i+1} - x}{h} f(x)\, dx.$$

On montre l'existence d'une solution du système (5.10).

THÉORÈME 5.1. *La matrice du système* (5.10) *est définie positive.*

DÉMONSTRATION. Soit Y un vecteur non nul de composantes y_i, $i = 1, \ldots, N-1$, auquel cas

$$Y^T AY = \sum_{j=1}^{N-1} \sum_{i=1}^{N-1} y_i [\varphi_i, \varphi_j]\, y_j.$$

On pose

$$\tilde{y}(x) = \sum_{i=1}^{N-1} y_i\, \varphi_i(x).$$

Alors

$$Y^T AY = [\tilde{y}, \tilde{y}] \geq \int_0^1 \left(\frac{d\tilde{y}}{dx}(x)\right)^2 dx.$$

L'inégalité de Cauchy-Bouniakovski aidant, on trouve

$$\int_0^x \left(\frac{d\tilde{y}(t)}{dt}\right)^2 dt \geq \left(\int_0^x \frac{d\tilde{y}(t)}{dt}\, dt\right)^2 .$$

Comme $\tilde{y}(0) = 0$, le second membre de la dernière inégalité vaut $(\tilde{y}(x))^2$. Les deux dernières relations se condensent en

$$Y^T A Y \geq (\tilde{y}(t))^2, \qquad t \in (0, 1),$$

et, en particulier,

$$Y^T A Y \geq y_i^2, \; i = 1, \ldots, N-1,$$

parce que $\tilde{y}(ih) = y_i$. D'où la minoration

$$Y^T A Y \geq h \sum_{i=1}^{N-1} y_i^2$$

qui garantit la définie positivité et la non-dégénérescence de la matrice A.

Ainsi, on associe à tout N un seul ensemble de constantes α_i tel que la solution approchée soit représentée par la formule (5.7).

THÉORÈME 5.2. *Si les coefficients du problème* (5.1), (5.2) *remplissent les conditions* (5.3), *la solution du problème approché* (5.7) *à* (5.9) *admet le développement*

$$u^h(x) = u(x) + \sum_{j=1}^{l} h^{2j} v_j(x) + h^{r+2} \eta^h(x), \qquad x \in \bar{\omega}_h. \quad (5.11)$$

Ici $l = [(r+1)/2]$, *les fonctions* v_j *sont de classe* $C^{r+4-2j}[0, 1]$ *et ne dépendent pas de* h, *et la fonction discrète* η^h *est bornée :*

$$|\eta^h(x)| \leq c_2 \qquad \forall\, x \in \bar{\omega}_h, \quad (5.12)$$

la constante c_2 *étant indépendante de* h.

DÉMONSTRATION. On note qu'il correspond à toute fonction $w(x)$ définie aux nœuds du réseau $\bar{\omega}_h$ une fonction continue

$$\tilde{w}(x) = \sum_{j=0}^{N} w(x_j)\, \varphi_i(x)$$

donnée sur $[0, 1]$ qui constitue *un prolongement linéaire par morceaux* de $w(x)$. Par exemple, $\tilde{u}^h = u^h(x)$. Le développement (5.11) se ramène donc à

$$u^h(x) = \tilde{u}(x) + \sum_{j=l} h^{2j} \tilde{v}_j(x) + h^{r+2} \tilde{\eta}^h(x), \qquad x \in [0, 1].$$

On pose

$$v_0 = u, \tag{5.13}$$

et on cherche v_j suivants sous forme de solutions des problèmes différentiels

$$\begin{aligned} -v_j'' + q(x)\,v_j &= R_j(x), & x &\in (0,\ 1), \\ v_j(0) = v_j(1) &= 0, & j &= 1,\ \ldots,\ l, \end{aligned} \tag{5.14}$$

où

$$R_1(x) = -\frac{1}{24}\,q(x)\,v_0''(x),$$

$$R_j(x) = -\sum_{s=1}^{j-1} \frac{2}{(2s+2)!}\,R_{j-s}^{(2s)}(x) -$$

$$-\sum_{s=1}^{j} \sum_{\substack{k+m=2s \\ k\geqslant 2,\ m\geqslant 0}} \frac{(m+1)(k^2+2km+3k-2m-4)}{(m+3)!\,k!}\,\frac{v_{j-s}^{(k)}(x)\,q^{(m)}(x)}{(k+m+1)(k+m+2)}, \tag{5.15}$$

$$j = 2,\ \ldots,\ l.$$

Ces formules permettent de constater que $R_1 \in C^r[0,\ 1]$, si bien qu'en vertu du théorème 1.1 le problème (5.14) possède pour $j = 1$ une solution unique dans $C^{r+2}\,[0,\ 1]$. On suppose $R_1,\ \ldots,\ R_j$ définies et telles que $R_k \in C^{r+2-2k}\,[0,\ 1]$. On trouve $v_1,\ \ldots,\ v_j$ comme solutions des problèmes (5.14) qui admettent, par le théorème 1.1, deux dérivées continues de plus que $R_1,\ \ldots,\ R_j$, i.e.

$$v_k \in C^{r+4-2k}\,[0,\ 1], \qquad k = 1,\ \ldots,\ j.$$

On établit, pour les termes du second membre de (5.15) de numéro $j+1$, le nombre maximal de dérivées:

$$R_{j+1-s}^{(2s)} \in C^{r-2j}\,[0,\ 1],$$

$$v_{j+1-s}^{(k)} \in C^{r+2-2j}\,[0,\ 1], \qquad s \geqslant 1,$$

$$q^{(m)} \in C^{r-2j}\,[0,\ 1]$$

(la dernière appartenance est juste parce que $m \leqslant j$). Ainsi, $R_{j+1} \in$ $\in C^{r-2j}\,[0,\ 1]$. D'après le théorème 1.1, le problème aux limites (5.14) de numéro $j+1$ admet une solution unique qui est de classe $C^{r+2-2j}\,[0,\ 1]$. C'est pourquoi toutes les fonctions R_j et v_j sont définies de façon unique, et

$$v_j \in C^{r-2j+4}\,[0,\ 1], \qquad R_j \in C^{r-2j+2}\,[0,\ 1]. \tag{5.16}$$

Connaissant v_j, u et u^h, on définit la fonction discrète

$$\eta^h(x) = \frac{1}{h^{r+2}}\left(u^h(x) - \sum_{j=0}^{l} v_j(x)\right), \qquad x \in \bar{\omega}_h. \tag{5.17}$$

ce qui équivaut à

$$\tilde{\eta}^h(x) = \frac{1}{h^{r+2}}\left(u^h(x) - \sum_{j=0}^{l} h^{2j}\tilde{v}_j(x)\right), \qquad x \in [0, 1]. \tag{5.18}$$

Conformément à (5.18), l'égalité (5.8)

$$[u^h, \varphi_i] = (f, \varphi_i), \qquad i = 1, \ldots, N-1,$$

conduit aux relations

$$[u^h, \varphi_i] = \sum_{j=0}^{l} h^{2j}[\tilde{v}_j, \varphi_i] + h^{r+2}[\tilde{\eta}^h, \varphi_i] = (f, \varphi_i), \tag{5.19}$$

$$i = 1, \ldots, N-1.$$

On se propose de les simplifier, et on établit au préalable plusieurs développements auxiliaires.

LEMME 5.3. *Soit t un entier naturel et v une fonction de $C^t[0, 1]$. On a*

$$(v, \varphi_i) = hv(x_i) + \sum_{j=1}^{\left[\frac{t-1}{2}\right]} h^{2j+1}\frac{2}{(2j+2)!}v^{(2j)}(x_i) + h^{t+1}\sigma_t, \tag{5.20}$$

où

$$|\sigma_t| \leqslant \frac{2}{(t+2)!}\max_{x\in[0,1]}|v^{(t)}(x)|.$$

DÉMONSTRATION. On suppose $g(x)$ telle que $g''(x) = v(x)$. Il suffit de poser par exemple

$$g(x) = \int_0^x\int_0^t v(z)\,dz\,dt,$$

et on vérifie sans peine la validité de la relation

$$(g'', \varphi_i) = \frac{g(x_{i-1}) - 2g(x_i) + g(x_{i+1})}{h}. \tag{5.21}$$

Le lemme 1.2, § 7.1 entraîne

$$(g'', \varphi_i) = h(g_{\bar{x}}(x_i))_{\hat{x}} =$$

$$= \sum_{j=0}^{[(t-1)/2]} h^{2j+1}\frac{2}{(2j+2)!}g^{(2j+2)}(x_i) + \frac{2h^{t+1}}{(t+2)!}g^{(t+2)}(\xi_i).$$

Comme $g^{(k+2)} = v^{(k)}$, on a le résultat voulu.

LEMME 5.4. *Etant donné $v \in C^t[0, 1]$, avec t un entier tel que $1 \leqslant t \leqslant r+2$, le prolongement linéaire par morceaux $\tilde{v}$ admet le développement*

$$(q\tilde{v}, \varphi_i) = (q\bar{v}, \varphi_i) + \\ + \sum_{j=1}^{[(t-1)/2]} h^{2j+1} \sum_{\substack{k+m=2j \\ k\geqslant 2, m\geqslant 0}} \frac{(m+1)(k^2+2km+3k-2m-4)}{(m+3)!\,k!} v^{(k)}(x_i)\, q^{(m)}(x_i) + \\ + h^{t+1}\varkappa_i^h, \tag{5.22}$$

où

$$|\varkappa_i^h| \leqslant c_3, \tag{5.23}$$

et c_3 est indépendante de i et h.

DÉMONSTRATION. On transforme le premier membre de (5.22). Le support de φ_i est concentré sur $[x_{i-1}, x_{i+1}]$, si bien que

$$(q\tilde{v}, \varphi_i) = \int_{x_{i-1}}^{x_{i+1}} q\tilde{v}\varphi_i\,dx.$$

Les fonctions $\tilde{v}$ et φ_i étant polynomiales par morceaux, il y a intérêt à partager l'intervalle d'intégration en deux intervalles partiels. Soit d'abord

$$\int_{x_i}^{x_{i+1}} q\tilde{v}\varphi_i\,dx = \frac{1}{h^2}\int_{x_i}^{x_{i+1}} q(x)\big(v(x_i)(x_{i+1}-x) + \\ + v(x_{i+1})(x-x_i)\big)(x_{i+1}-x)\,dx.$$

On remplace $v(x_{i+1})$ par son développement taylorien autour du point x_i, il vient

$$\int_{x_i}^{x_{i+1}} q\tilde{v}\varphi_i\,dx = \frac{1}{h}\int_{x_i}^{x_{i+1}} q(x)\,v(x_i)(x_{i+1}-x)\,dx + \\ + \frac{1}{h}\int_{x_i}^{x_{i+1}} q(x)\,v'(x_i)(x-x_i)(x_{i+1}-x)\,dx + \\ + \sum_{k=2}^{r+1} \frac{h^{k-2}}{k!} v^{(k)}(x_i) \int_{x_i}^{x_{i+1}} (x-x_i)(x_{i+1}-x)\,q(x)\,dx + \\ + \frac{h^{r+3}}{(r+2)!} v^{(r+2)}(\xi_i)\,\nu_i^h, \tag{5.24}$$

où

$$|\nu_i^k| \leqslant \frac{1}{6} \max_{[0,1]} |q|.$$

On substitue à $q(x)$ sous le signe somme son développement taylorien par rapport au point x_i. On a après avoir intégré :

$$\frac{h^{k-2}}{k!} v^{(k)}(x_i) \int_{x_i}^{x_{i+1}} (x - x_i)(x_{i+1} - x) \times$$

$$\times \left[\sum_{m=0}^{r+1-k} \frac{(x - x_i)^m}{m!} q^{(m)}(x_i) + \frac{(x - x_i)^{r+2-k}}{(r+2-k)!} q^{(r+2-k)}(\xi_i) \right] dx =$$

$$= \sum_{m=0}^{r+1-k} \frac{h^{k+m+1}(m+1)}{k!(m+3)!} v^{(k)}(x_i) q^{(m)}(x_i) + h^{r+3} \mu_{k,i}^h,$$

où

$$|\mu_{k,i}^h| \leqslant \frac{r+3-k}{k!(r+5-k)!} \max_{[0,1]} |q^{(r+2-k)}| \max_{[0,1]} |v^{(k)}|.$$

Finalement,

$$\int_{x_i}^{x_{i+1}} q\, \sigma\, \varphi_i\, dx = \frac{1}{h} \int_{x_i}^{x_{i+1}} q(x) v(x_i)(x_{i+1} - x)\, dx +$$

$$+ \frac{1}{h} \int_{x_i}^{x_{i+1}} q(x) v'(x_i)(x_{i+1} - x)(x - x_i)\, dx +$$

$$+ \sum_{k=2}^{r+1} \sum_{m=0}^{r+1-k} \frac{h^{k+m+1}(m+1)}{(m+3)!\,k!} v^{(k)}(x_i) q^{(m)}(x_i) + h^{r+3} \mu_i^h \qquad (5.25)$$

où

$$|\mu_i^h| \leqslant c_4, \qquad i = 1, \ldots, N-1.$$

On traite de même l'intégrale

$$\int_{x_i}^{x_{i+1}} qv\, \varphi_i\, dx.$$

et on aboutit à la formule

$$\int_{x_i}^{x_{i+1}} q\, v\, \varphi_i\, dx = \frac{1}{h} \int_{x_i}^{x_{i+1}} q(x)\, v(x_i)\, (x_{i+1} - x)\, dx +$$

$$+ \frac{1}{h} \int_{x_i}^{x_{i+1}} q(x)\, v'(x_i)\, (x_{i+1} - x)\, (x - x_i)\, dx +$$

$$+ \sum_{k=2}^{r+1} \sum_{m=0}^{r+1-k} \frac{h^{k+m+1}}{m!\, k!\, (k+m+1)\, (k+m+2)} v^{(k)}(x_i)\, q^{(m)}(x_i) + h^{r+3} \rho_i^h, \quad (5.26)$$

où

$$|\rho_i^h| \leqslant c_5, \qquad i = 1, \ldots, N-1.$$

On fait la différence de (5.25) et (5.26):

$$\int_{x_i}^{x_{i+1}} q\, \tilde{v}\, \varphi_i\, dx - \int_{x_i}^{x_{i+1}} q\, v\, \varphi_i\, dx =$$

$$= \sum_{k=2}^{r+1} \sum_{m=0}^{r+1-k} \frac{h^{k+m+1} (m+1)\, (k^2 + 2km + 3k - 2m - 4)}{(m+3)!\, k!\, (k+m+2)\, (k+m+1)} \times$$

$$\times v^{(k)}(x_i)\, q^{(m)}(x_i) + h^{r+3} (\mu_i^h - \rho_i^h). \quad (5.27)$$

Si l'on raisonne non sur x_{i+1}, mais sur x_{i-1} (et on remplace donc h du second membre de (5.27) par $-h$) et si l'on intervertit l'ordre d'intégration (si bien que le second membre de (5.27) change de signe), alors

$$\int_{x_{i-1}}^{x_i} q\, \tilde{v}\, \varphi_i\, dx - \int_{x_{i-1}}^{x_i} q\, v\, \varphi_i\, dx =$$

$$= -\sum_{k=2}^{r+1} \sum_{m=0}^{r+1-k} \frac{(-h)^{k+m+1} (m+1)\, (k^2 + 2km + 3k - 2m - 4)}{(m+3)!\, k!\, (k+m+1)\, (k+m+2)} \times$$

$$\times v^{(k)}(x_i)\, q^{(m)}(x_i) + h^{r+3} \delta_i^h.$$

où

$$|\delta_i^h| \leqslant c_4 + c_5.$$

On additionne (5.27) et la dernière égalité, ce qui donne l'affirmation du lemme 5.4, où $c_3 = 2c_4 + 2c_5$.

On note que c'est la nature du problème qui a dicté la démonstration. En effet, la fonction q n'est pas suffisamment régulière pour qu'on puisse procéder de façon plus simple et développer v et q à la fois.

LEMME 5.5. *On a pour toute fonction $v \in C^1\ [0, 1]$ et son prolongement linéaire par morceaux*

$$(v', \varphi_i') = (\tilde{v}', \varphi_i').$$

DÉMONSTRATION. Le premier membre devient par certaines transformations

$$(v', \varphi_i') = -\frac{1}{h}\int_{x_i}^{x_{i+1}} v'(x)\,dx + \frac{1}{h}\int_{x_{i-1}}^{x_i} v'(x)\,dx =$$

$$= -\frac{1}{h}\left[v(x_{i+1}) - v(x_i)\right] + \frac{1}{h}\left[v(x_i) - v(x_{i+1})\right]. \quad (5.28)$$

Comme

$$\tilde{v}'(x) = \frac{v(x_{i+1}) - v(x_i)}{h} \quad \text{pour} \quad x \in [x_i, x_{i+1}],$$

$$\tilde{v}'(x) = \frac{v(x_i) - v(x_{i-1})}{h} \quad \text{pour} \quad x \in [x_{i-1}, x_i],$$

on a

$$(\tilde{v}', \varphi_i') = -\frac{1}{h}\int_{x_i}^{x_{i+1}} \frac{v(x_{i+1}) - v(x_i)}{h}\,dx + \frac{1}{h}\int_{x_{i-1}}^{x_i} \frac{v(x_i) - v(x_{i-1})}{h}\,dx.$$

On démontre le lemme par identification avec (5.28).

Soit de nouveau l'égalité (5.19). On récrit son premier membre par recours au lemme 5.5:

$$[u^h, \varphi_i] = \sum_{j=0}^{l} h^{2j}\{(\tilde{v}_j', \varphi_i') + (q\,\tilde{v}_j, \varphi_i)\} + h^{r+2}[\eta^h, \varphi_i] =$$

$$= \sum_{j=0}^{l} h^{2j}\{(v_j', \varphi_i') + (q\,\tilde{v}_j, \varphi_i)\} + h^{r+2}\ [\eta^h, \varphi_i].$$

On transforme certains termes par le lemme 5.4:

$$[u^h, \varphi_i] = \sum_{j=0}^{l} h^{2j}\{(v_j', \varphi_i') + (qv_j, \varphi_i)\} +$$

$$+\sum_{j=0}^{l-1}\sum_{s=1}^{l-j} h^{2j+2s+1} \sum_{\substack{k+m=2s\\ k\geq 2,\ m\geq 0}} \frac{(m+1)(k^2+2km+3k-2m-4)\, v_j^{(k)}(x_i)\, q^{(m)}(x_i)}{(m+3)!\, k!\, (k+m+1)(k+m+2)} +$$

$$+ h^{r+3}\beta_i^h + h^{r+2}[\tilde{\eta}^h, \varphi_i], \qquad (5.29)$$

où

$$|\beta_i^h| \leq c_6.$$

Les fonctions v_j étant solutions des problèmes (5.14), (5.15), on a les égalités

$$[v_j, \varphi_i] = (v_j', \varphi_i') + (qv_j, \varphi_i) = (R_j, \varphi_i) \qquad (5.30)$$

obtenues en intégrant par parties l'équation (5.14) multipliée par φ_i. De plus, selon le lemme 5.3,

$$(R_j, \varphi_i) = \sum_{s=0}^{l-j} h^{2s+1}\frac{2}{(2s+2)!} R_j^{(2s)}(x_i) + h^{r+3-2j} S_{j,i}^h,$$

avec

$$|S_{j,i}^h| \leq c_7, \qquad i = 1, \ldots, N-1; \qquad j = 1, \ldots, l.$$

On transforme (5.29) à la lumière du dernier développement et des formules (5.30) et (5,8):

$$[u^h, \varphi_i] = \sum_{j=1}^{l} h^{2j}\sum_{s=0}^{l-j} h^{2s+1}\frac{2}{(2s+2)!} R_j^{(2s)}(x_i) +$$

$$+\sum_{j=0}^{l}\sum_{s=1}^{l-j} h^{2j+2s+1} \sum_{\substack{k+m=2s\\ k\geq 2,\ m\geq 0}} \frac{(m+1)(k^2+2km+3k-2m-4)\, v_j^{(k)}(x_i)\, q^{(m)}(x_i)}{(m+3)!\, k!\, (k+m+1)(k+m+2)} +$$

$$+ h^{r+3} T_i^h + h^{r+2}[\tilde{\eta}^h, \varphi_i] + (f, \varphi_i).$$

Ici

$$|T_i^h| \leq c_8, \qquad i = 1, \ldots, N-1.$$

On intervertit l'ordre de sommation dans les sommes doubles:

$$[u^h, \varphi_i] = \sum_{j=1}^{l} h^{2j+1} \left\{ \sum_{s=0}^{j-1} \frac{2}{(2s+2)!} R_{j-s}^{(2s)}(x_i) + \right.$$

$$\left. + \sum_{s=1}^{j} \sum_{\substack{k+m=2s \\ k\geq 2,\ m\geq 0}} \frac{(m+1)(k^2+2km+3k-2m-4)\, v_{j-s}^{(k)}(x_i)}{(m+3)!\,k!} \frac{q^{(m)}(x_i)}{(k+m+1)(k+m+2)} \right\} +$$

$$+ h^{r+3} T_i^h + h^{r+2} [\tilde{\eta}^h, \varphi_i] + (f, \varphi_i).$$

Il est évident qu'avec le choix des fonctions R_j (formule (5.15)), la somme double s'annule:

$$[u^h, \varphi_i] = h^{r+3} T_i^h + h^{r+2} [\tilde{\eta}^h, \varphi_i] + (f, \varphi_i).$$

On a par définition de u^h

$$[u^h, \varphi_i] = (f, \varphi_i),$$

si bien que

$$[\tilde{\eta}^h, \varphi_i] = -hT_i^h, \qquad i = 1, \ldots, N-1.$$

On fait le produit de chaque équation par $\eta^h(x_i)$ et on somme par rapport à tous les indices $i = 1, \ldots, N-1$. Vu les égalités $\tilde{\eta}^h(0) = \tilde{\eta}^h(1) = 0$ (résultant de la définition de $\tilde{\eta}^h$ aux points 0 et 1), on trouve

$$[\tilde{\eta}^h, \tilde{\eta}^h] = -\sum_{i=1}^{N-1} T_i^h \eta^h(x_i)\, h.$$

On a montré (théorème 5.1) que

$$(\tilde{\eta}^h(x))^2 \leqslant [\tilde{\eta}^h, \tilde{\eta}^h] \qquad \forall x \in \omega_h;$$

donc

$$\max_{x\in\omega_h} |\tilde{\eta}^h(x)|^2 \leqslant \sum_{i=1}^{N-1} |T_i^h|\, |\eta^h(x_i)|\, h \leqslant \max_{x\in\omega_h} |\tilde{\eta}^h(x)|\, c_8.$$

On simplifie par $\max\limits_{x\in\omega_h} |\tilde{\eta}^h(x)|$:

$$\max_{x\in\omega_h} |\tilde{\eta}^h(x)| \leqslant c_8,$$

ce qui démontre le théorème 5.2, où $c^2 = c_8$.

Ci-dessous un algorithme d'extrapolation basé sur le développement du théorème 5.2.

Soit $l = [(r+1)/2]$, avec r un entier de la condition (5.3). On fait correspondre à $l+1$ entiers $N_k = kN$ les réseaux $\bar{\omega}_{h_k}$ de pas $h_k = 1/N_k$, $k = 1, \ldots, l+1$. On énonce pour chaque $\bar{\omega}_{h_k}$ le problème (5.8), (5.9) et on cherche sa solution avec les poids α_i pris dans (5.7). On trouve $l+1$ fonctions $u^{h_k}(x)$ dont les valeurs aux nœuds de $\bar{\omega}_{h_1}$ nous intéresseront seules. On forme la combinaison linéaire

$$U(x) = \sum_{k=0}^{l+1} \gamma_k u^{h_k}(x), \qquad x \in \bar{\omega}_{h_1}, \tag{5.31}$$

où les poids

$$\gamma_k = 2\frac{(-1)^{l-k+1} k^{2l+2}}{(l-k+1)!\,(l+k+1)!}. \tag{5.32}$$

On a le

THÉORÈME 5.6. *Hypothèses du théorème 5.2. Alors la solution corrigée* (5.31), (5.32) *admet l'estimation*

$$\max_{x \in \bar{\omega}_{h_1}} |U(x) - u(x)| \leqslant h_1^{r+2} c_9 \tag{5.33}$$

avec la constante c_9 indépendante de h_k.

La démonstration imite celle du théorème 4.6, et la constante c_9 est égale à

$$c_9 = 2c_2 \sum_{k=1}^{l+1} \frac{k^{2l-r}}{(l+1-k)!\,(l+1+k)!},$$

avec c_2 de la condition du théorème 5.2.

Soit l'équation plus générale

$$\begin{gathered} -(pu')' + qu = f \quad \text{sur} \quad (0, 1), \\ p(x) \geqslant c_1 > 0, \qquad q(x) \geqslant 0, \\ p \in C^{r+1}[0, 1], \qquad f, q \in C^r[0, 1]. \end{gathered} \tag{5.34}$$

On généralise les résultats ci-dessus de deux manières différentes.

Dans le premier procédé, les fonctions de base (5.4) ne varient pas, et l'on a le problème (5.7) à (5.10), où $[v, w]$ a le sens suivant:

$$[v, w] = \int_0^1 pv'w'dx + \int_0^1 qvw\,dx. \tag{5.35}$$

Le système d'équations algébriques (5.10) est toujours de structure matricielle tridiagonale, mais les coefficients a_i, b_i, c_i, g_i ne sont plus les mêmes:

$$a_i = [\varphi_{i-1}, \varphi_i] = -\frac{1}{h^2}\int_{x_{i-1}}^{x_i} p(x)\,dx + \int_{x_{i-1}}^{x_i} \frac{x_i - x}{h}\,\frac{x - x_{i-1}}{h}\,q(x)\,dx,$$

$$c_i = a_{i+1} = [\varphi_i, \varphi_{i+1}],$$

$$b_i = [\varphi_i, \varphi_i] = \frac{1}{h^2}\int_{x_{i-1}}^{x_i} p(x)\,dx + \frac{1}{h^2}\int_{x_i}^{x_{i+1}} p(x)\,dx +$$

$$+ \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1})^2}{h^2}\,q(x)\,dx + \int_{x_i}^{x_{i+1}} \frac{(x_{i+1} - x)^2}{h^2}\,q(x)\,dx, \tag{5.36}$$

$$g_i = (f, \varphi_i) = \int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{h}\,f(x)\,dx + \int_{x_i}^{x_{i+1}} \frac{x_{i+1} - x}{h}\,f(x)\,dx.$$

Les lemmes 5.3 et 5.4 restent valides tandis que le lemme 5.5 n'a plus lieu. On le remplace par le développement de la différence $(p\bar{v}', \varphi_i') - (pv', \varphi_i')$ par rapport à h, ce qui donne un reste en h^{r+2}. Finalement, le reste de (5.11) sera de l'ordre de h^{r+1} sous les mêmes conditions imposées à q et f.

Le second procédé consiste à utiliser les fonctions de base φ_i plus compliqués *, à savoir

$$\varphi_i(x) = \begin{cases} \int_{x_{i-1}}^{x} p^{-1}(t)\,dt \Big/ \int_{x_{i-1}}^{x_i} p^{-1}(t)\,dt & \text{si} \quad x - x_i \in (-h, 0], \\ \int_{x}^{x_{i+1}} p^{-1}(t)\,dt \Big/ \int_{x_i}^{x_{i+1}} p^{-1}(t)\,dt & \text{si} \quad x - x_i \in (0, h], \\ 0 & \text{dans les cas restants.} \end{cases} \tag{5.37}$$

* On trouve des fonctions analogues dans [1], [47], [115].

Les coefficients du système (5.10) se compliquent eux aussi :

$$a_i = [\varphi_{i-1}, \varphi_i] = \frac{-1}{\int_{x_{i-1}}^{x_i} p^{-1}(t)\,dt} + \int_{x_{i-1}}^{x_i} \frac{\int_x^{x_i} p^{-1}(t)\,dt \int_{x_{i-1}}^{x} p^{-1}(t)\,dt}{\left(\int_{x_{i-1}}^{x_i} p^{-1}(t)\,dt\right)^2}\, q(x)\,dx,$$

$$c_i = a_{i+1} = [\varphi_i, \varphi_{i+1}],$$

$$b_i = \frac{1}{\int_{x_{i-1}}^{x_i} p^{-1}(t)\,dt} + \frac{1}{\int_{x_i}^{x_{i+1}} p^{-1}(t)\,dt} + \frac{\int_{x_{i-1}}^{x_i} \left(\int_{x_{i-1}}^{x} p^{-1}(t)\,dt\right)^2 q(x)\,dx}{\left(\int_{x_{i-1}}^{x_i} p^{-1}(t)\,dt\right)^2} + \frac{\int_{x_i}^{x_{i+1}} \left(\int_{x}^{x_{i+1}} p^{-1}(t)\,dt\right)^2 q(x)\,dx}{\left(\int_{x_i}^{x_{i+1}} p^{-1}(t)\,dt\right)^2}, \qquad (5.38)$$

$$g_i = \frac{\int_{x_{i-1}}^{x_i} \left(\int_{x_{i-1}}^{x} p^{-1}(t)\,dt\right) f(x)\,dx}{\int_{x_{i-1}}^{x_i} p^{-1}(t)\,dt} + \frac{\int_{x_i}^{x_{i+1}} \left(\int_{x}^{x_{i+1}} p^{-1}(t)\,dt\right) f(x)\,dx}{\int_{x_i}^{x_{i+1}} p^{-1}(t)\,dt}.$$

S'agissant de ces fonctions d'essai, le lemme 5.3 et l'égalité $(pv', \varphi_i') = (p\tilde{v}', \varphi_i')$ restent justes, ainsi que d'ailleurs le lemme 5.4 dont la démonstration devient sensiblement plus ardue encore qu'elle donne un reste de même ordre de grandeur.

Fort des résultats obtenus, on dit donc qu'avec les fonctions d'essai (5.37), on conserve l'ordre du reste du développement (5.11). On note que la dernière égalité signifie pour $q \equiv 0$ l'intégration exacte de l'équation (5.34) dans la méthode décrite.

On observe qu'en ce qui concerne h réels, la précision réalisée avec les fonctions de base (5.37) est supérieure à celle obtenue avec

les fonctions (5.4). La chose est particulièrement évidente dans les problèmes où p est à variation forte.

On l'illustre par le problème

$$-\frac{d}{dx}\left(e^{-4x}\frac{du}{dx}\right)+4e^{-4x}u=8-4x+4xe^{-4x},\qquad u(0)=u(1)=0. \tag{5.39}$$

Sa solution exacte est la fonction

$$u(x)=(e^{4x}-1)(1-x).$$

On a résolu plusieurs problèmes variationnels aux différences (5.10) avec les fonctions d'essai simples (5.4). Dans ce cas, les coef-

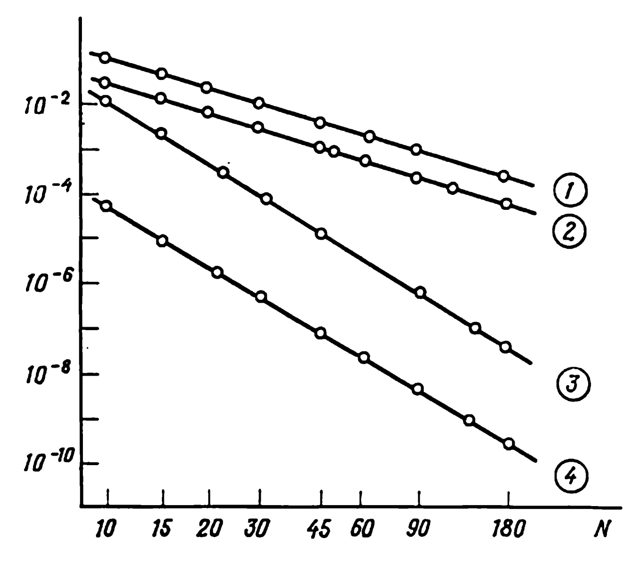


Fig. 3.3. Erreurs maxima sur les solutions variationnelles aux différences et sur les fonctions extrapolées aux nœuds des réseaux de discrétisation

Erreur sur la solution variationnelle aux différences: *1* — cas des fonctions de base linéaires par morceaux; *2* — cas des fonctions de base spéciales (5.37). Erreur sur les solutions extrapolées: *3* — cas des fonctions de base linéaires par morceaux; *4* — cas des fonctions de base (5.37).

ficients du système (5.10) sont définis par les formules (5.36). On a trouvé l'écart maximum entre chaque solution approchée (5.7) et la solution correcte. Le graphe correspondant est représenté sur la fig. 3.3. On a ensuite abordé les problèmes (5.10) avec les fonctions d'essai (5.37), auquel cas les coefficients du système sont définis par les formules (5.38). Connaissant les poids, on a formé les solutions

approchées (5.7) et établi dans chaque fois l'écart maximum. On voit le graphe de cette quantité sur la fig. 3.3. Afin d'apprécier l'amélioration apportée par l'extrapolation, on a formé les solutions corrigées (5.31) pour $l = 1$, i.e. on a extrapolé sur les valeurs de deux solutions des problèmes approchés (5.10) pour le rapport de pas 2 : 1. On a appliqué l'extrapolation aussi bien aux solutions construites à l'aide des fonctions d'essai (5.4) qu'à celles obtenues moyennant (5.37), et cela sur deux réseaux de pas $h = 1/N$ et $h/2$. Pour les erreurs maxima sur les solutions améliorées, voir fig. 3.3.

3.6. Equation quasi linéaire

Considérons le problème de Dirichlet relatif à une équation quasi linéaire du second ordre et supposons que l'équation soit résolue explicitement par rapport à la dérivée seconde et que le second membre soit une fonction d'une variable indépendante, de la solution et d'une dérivée de la solution. Pour qu'il y ait existence et unicité pour un problème non linéaire, il faut (c'est bien connu) que le second membre soit suffisamment régulier et qu'on soit dans certaines conditions supplémentaires. Dans ce paragraphe, on démontrera un théorème aux termes duquel l'analogue discret d'un problème non linéaire est représentable sous forme de développement suivant les puissances entières du pas du réseau. Ce résultat est le pivot de la construction d'une solution améliorée par l'extrapolation de Richardson, et avec un jeu de solutions approchées associées à divers paramètres de discrétisation, on atteint, comme dans le cas linéaire, une précision maximale.

Soit le problème

$$u'' = f(x, u, u') \quad \text{sur} \quad [0, 1], \tag{6.1}$$

$$u(0) = u_0, \qquad u(1) = u_1. \tag{6.2}$$

On suppose que $f(x, u, v)$ appartient à $C^r\,([0, 1] \times (-\infty, \infty) \times (-\infty, \infty))$, $r \geq 2$, en tant que fonctions de trois variables et que le problème proposé admet une solution de classe $C^{r+2}\,[0, 1]$.

Comme l'extrapolation de Richardson pour les équations non linéaires ne sera décrite que sommairement, nous bornons à un seul critère d'existence et d'unicité, disons, à

$$\frac{1}{2} - \frac{1}{16}\left|\frac{\partial f}{\partial v}(x, u, v)\right|^2 + \frac{1}{8}\,\frac{\partial f}{\partial u}(x, u, v) \geq c_1, \qquad c_1 \in \left(0, \frac{1}{2}\right), \tag{6.3}$$
$$x \in [0, 1], \qquad u \in (-\infty, \infty), \qquad v \in (-\infty, \infty).$$

On aurait tort de se passer de tout critère et de supposer tout bonnement l'existence et l'unicité de la solution. En effet, les critères de possibilité du problème différentiel déterminent des fois le choix

du schéma aux différences en en induisant la possibilité, l'unicité et la stabilité.

On montre que la condition (6.3) implique l'inégalité caractéristique de la monotonie forte (voir [26], [38]) du problème (6.1), (6.2)

$$\int_0^1 \{u'(u'-v')+f(x,u,u')(u-v)\}\,dx -$$

$$-\int_0^1 \{v'(u'-v')+f(x,v,v')(u-v)\}\,dx \geq c_1\int_0^1 (u'-v')^2\,dx \quad (6.4)$$

pour toute u et toute v de C^1 $[0, 1]$ qui sont égales aux extrémités du segment: $u(0)=v(0)$, $u(1)=v(1)$.

On pose

$$q_0=\tau u+(1-\tau)v, \qquad q_1=\tau u'+(1-\tau)v',$$

auquel cas

$$f(x,u,u')-f(x,v,v')=\int_0^1 \frac{d}{d\tau}f(x,q_0,q_1)\,d\tau=$$

$$=\int_0^1 \frac{\partial f}{\partial q_0}\,d\tau\,(u-v)+\int_0^1 \frac{\partial f}{\partial q_1}\,d\tau\,(u'-v').$$

On désigne par $J(u,v)$ le premier membre de (6.4), il vient par suite de la relation précédente

$$J(u,v)=\int_0^1\left\{\int_0^1 \frac{\partial f}{\partial q_0}(x,q_0,q_1)\,d\tau\,(u-v)^2+\right.$$

$$\left.+\int_0^1 \frac{\partial f}{\partial q_1}(x,q_0,q_1)\,d\tau\,(u-v)(u'-v')+(u'-v')^2\right\}dx.$$

On retranche de deux membres $c_1\int_0^1 (u'-v')^2dx$ et on utilise l'inégalité $ab\leq a^2\varepsilon/2+b^2/(2\varepsilon)$, $\varepsilon>0$. On a

$$J(u,v)-c_1\int_0^1 (u'-v')^2\,dx\geq\int_0^1\left\{\int_0^1 \frac{\partial f}{\partial q_0}(x,q_0,q_1)\,d\tau\,(u-v)^2-\right.$$

$$-\int_0^1\left|\frac{\partial f}{\partial q_1}(x,q_0,q_1)\right|d\tau\left(\frac{\varepsilon}{2}(u-v)^2+\frac{1}{2\varepsilon}(u'-v')^2\right)+$$

$$\left.+(1-c_1)(u'-v')^2\right\}dx. \quad (6.5)$$

Les fonctions u et v prenant les mêmes valeurs aux extrémités du segment $[0, 1]$ vérifient l'inégalité (voir [43])

$$\frac{1}{8}\int_0^1 (u' - v')^2\, dx \geqslant \int_0^1 (u - v)^2\, dx.$$

On pose $\varepsilon = \int_0^1 \left| \frac{\partial f}{\partial q_1}(x, q_0, q_1) \right| d\tau$ et on récrit le second membre de (6.5) compte tenu de l'inégalité ci-dessus:

$$J(u, v) - c_1 \int_0^1 (u' - v')^2\, dx \geqslant \int_0^1 \left\{ \int_0^1 \left(4 - 8c_1 - \right.\right.$$

$$\left.\left. - \frac{1}{2} \left| \frac{\partial f}{\partial q_1}(x, q_0, q_1) \right|^2 + \frac{\partial f}{\partial q_0}(x, q_0, q_1) \right) d\tau \right\} (u - v)^2\, dx.$$

Le second membre est non négatif par l'inégalité (6.3). Aussi

$$J(u, v) \geqslant c_1 \int_0^1 (u' - v')^2\, dx,$$

ce qui est équivalent à (6.4).

Le problème étant fortement monotone admet une solution unique. Soit, par exemple, u et v deux solutions distinctes de classe $C^{r+2}[0, 1]$ du problème (6.1), (6.2). Dans ce cas, $u(0) = v(0)$, $u(1) = v(1)$ et on a l'inégalité (6.4) dont le premier membre est 0. Mais le théorème de l'immersion de $W_2^1[0, 1]$ dans $C[0, 1]$ (voir [43]) entraîne

$$\max_{[0,1]} |u - v| \leqslant \frac{1}{2} \left(\int_0^1 (u' - v')^2\, dx \right)^{1/2}.$$

D'où $u = v$ sur $[0, 1]$, ce qui contredit l'hypothèse de deux solutions.

La résolution numérique du problème (6.1), (6.2) utilise le schéma aux différences

$$u^h_{\bar{x}\hat{x}}(x) = f_{\hat{x}}\left(x, u^h_{\hat{x}}(x), u^h_{\hat{x}}(x)\right), \qquad x \in \omega_h, \tag{6.6}$$

$$u^h(0) = u_0, \quad u^h(1) = u_1. \tag{6.7}$$

Avec d'autres notations, l'équation (6.6) s'écrit

$$\frac{u(x-h) - 2u(x) + u(x+h)}{h^2} =$$

$$= \frac{1}{2} f\left(x + h/2, \frac{u(x+h) + u(x)}{2}, \frac{u(x+h) - u(x)}{h} \right) +$$

$$+ \frac{1}{2} f\left(x - h/2, \frac{u(x) + u(x-h)}{2}, \frac{u(x) - u(x-h)}{h} \right), \quad x \in \omega_h.$$

On démontre que le problème (6.6), (6.7) admet une solution unique. On montre au préalable qu'il jouit pour h suffisamment petits de la propriété discrète de monotonie forte:

$$\sum_{x\in\omega_h} \{-u_{\bar{x}\hat{x}}(x) + f_{\bar{x}}(x, u_{\bar{x}}(x), u_{\hat{x}}(x))\}(u(x) - v(x))\,h -$$
$$-\sum_{x\in\omega_h} \{-v_{\bar{x}\hat{x}}(x) + f_{\bar{x}}(x, v_{\bar{x}}(x), v_{\hat{x}}(x))\}(u(x) - v(x))\,h \geqslant$$
$$\geqslant c_1 \sum_{x\in\bar{\omega}_h} (u_{\hat{x}}(x) - v_{\hat{x}}(x))^2 h. \qquad (6.8)$$

avec u et v des fonctions quelconques définies sur ω_h telles que $u(0) = v(0)$ et $u(1) = v(1)$. Comme dans le cas différentiel on pose

$$q_0(x) = \tau u_{\bar{x}}(x) + (1-\tau)\,v_{\bar{x}}(x), \qquad x\in\bar{\omega}_h,$$
$$q_1(x) = \tau u_{\hat{x}}(x) + (1-\tau)\,v_{\hat{x}}(x), \qquad x\in\bar{\omega}_h.$$

Alors

$$f(x, u_{\bar{x}}, u_{\hat{x}}) - f(x, v_{\bar{x}}, v_{\hat{x}}) = \int_0^1 \frac{d}{d\tau} f(x, q_0, q_1)\,d\tau =$$
$$= \int_0^1 \frac{\partial f}{\partial q_0}\,d\tau\,(u_{\bar{x}} - v_{\bar{x}}) + \int_0^1 \frac{\partial f}{\partial q_1}\,d\tau\,(u_{\hat{x}} - v_{\hat{x}}) \qquad \forall x\in\bar{\omega}_h.$$

On applique à la différence divisée seconde du premier membre de (6.8) la formule de Green, et on transforme les autres termes par la dernière identité, il vient

$$J^h(u, v) = \sum_{x\in\bar{\omega}_h} \Big\{(u_{\hat{x}}(x) - v_{\hat{x}}(x))^2 +$$
$$+ \int_0^1 \frac{\partial f(x, q_0, q_1)}{\partial q_0}\,d\tau\,(u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 + \int_0^1 \frac{\partial f(x, q_0, q_1)}{\partial q_1}\,d\tau \times$$
$$\times (u_{\hat{x}}(x) - v_{\hat{x}}(x))(u_{\bar{x}}(x) - v_{\bar{x}}(x))\Big\}\,h.$$

On a utilisé la propriété de u et v de coïncider aux extrémités de $[0, 1]$ et l'identité aux différences

$$\sum_{x\in\omega_h} w_{\bar{x}}(x)(u(x) - v(x)) = \sum_{x\in\bar{\omega}_h} w(x)(u_{\hat{x}}(x) - v_{\hat{x}}(x))$$

juste pour toute fonction w définie aux nœuds de $\bar{\omega}_h$. On effectue sur $J^h(u, v)$ des transformations analogues à celles du cas continu et on a

$$J^h(u, v) - c_1 \sum_{x \in \bar{\omega}_h} (u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 h \geqslant$$

$$\geqslant \sum_{x \in \bar{\omega}_h} \left\{ \int_0^1 \frac{\partial f}{\partial q_0}(x, q_0, q_1)\, d\tau \,(u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 - \right.$$

$$- \frac{1}{2} \int_0^1 \left| \frac{\partial f}{\partial q_1}(x, q_0, q_1) \right|^2 d\tau \,(u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 +$$

$$\left. + \left(\frac{1}{2} - c_1 \right) (u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 \right\} h.$$

On écrit deux inégalités vérifiées par w définie sur $\bar{\omega}_h$ et nulle pour $x = 0$ et $x = 1$:

$$\sum_{x \in \bar{\omega}_h} (w_{\bar{x}}(x))^2 h \geqslant 8 \sum_{x \in \bar{\omega}_h} w^2(x) h.$$

$$\sum_{x \in \omega_h} w^2(x) h \geqslant \sum_{x \in \bar{\omega}_h} (w_{\bar{x}}(x))^2 h.$$

La première est un analogue aux différences de l'estimation en norme de l'immersion de $\overset{\circ}{W}{}_2^1[0, 1]$ dans $L_2[0, 1]$ qu'on prouve dans [43], et la seconde découle presque évidemment de la relation

$$w(x)\, w(x + h) \leqslant (w^2(x) + w^2(x + h))/2.$$

Avec ces deux inégalités, on trouve

$$J^h(u, v) - c_1 \sum_{x \in \bar{\omega}_h} (u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 h \geqslant$$

$$\geqslant \sum_{x \in \bar{\omega}_h} \left\{ \int_0^1 \left(4 - 8c_1 - \frac{1}{2} \left| \frac{\partial f}{\partial q_1}(x, q_0, q_1) \right|^2 + \right. \right.$$

$$\left. \left. + \frac{\partial f}{\partial q_0}(x, q_0, q_1) \right) d\tau \right\} (u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 h.$$

Comme le second membre est non négatif selon (6.3), on a

$$J^h(u, v) \geqslant c_1 \sum_{x \in \bar{\omega}_h} (u_{\bar{x}}(x) - v_{\bar{x}}(x))^2 h,$$

ce qui est équivalent à (6.8) par suite de la notation $J^h(u, v)$.

On utilise l'analogue discret de la propriété de monotonie forte en passant des conditions aux limites (6.7) aux conditions homogènes. On introduit donc la fonction linéaire

$$F(x) = u_0 + x(u_1 - u_0)$$

et on effectue la substitution

$$u^h(x) = F(x) + \xi(x), \qquad x \in \omega_h,$$

auquel cas ξ satisfait au problème

$$\begin{gathered} \xi_{\bar{x}x}(x) = f_{\bar{x}}(x, \ \xi_{\bar{x}}(x) + F_{\bar{x}}(x), \ \xi_{\dot{x}}(x) + F_{\dot{x}}(x)), \quad x \in \omega_h, \\ \xi(0) = \xi(1) = 0, \end{gathered} \tag{6.9}$$

qui est possible s'il en est de même du problème (6.6), (6.7).

Les raisonnements suivants imitent ceux de [38]. Ils s'inspirent d'un lemme de [111] dont voici l'énoncé.

LEMME 6.1. *Soit $\xi \to P(\xi)$ une application continue de E^m dans lui-meme telle que $(P(\xi), \xi) \geqslant 0$ pour un certain $\rho > 0$ et tout ξ de la sphère $\|\xi\| = \rho$. Il existe ξ, $\|\xi\| \leqslant \rho$, pour lequel $P(\xi) = 0$.*

Pour qu'il soit possible d'utiliser ce résultat, on prend une application de l'espace de dimension $N - 1$ dans lui-même telle qu'il corresponde à chaque vecteur ξ de composantes $\xi_i = \xi(x_i)$, $x_i \in \omega_h$, un vecteur P_ξ de composantes

$$P_\xi(x_i) = -\xi_{\bar{x}\dot{x}}(x_i) + f_{\bar{x}}(x_i, \ \xi_{\bar{x}}(x_i) + F_{\bar{x}}(x_i), \ \xi_{\dot{x}}(x_i) + F_{\dot{x}}(x_i)).$$

On a unifié les notations en posant $\xi(0) = 0$ et $\xi(1) = 0$. La continuité de l'application $\xi \to P_\xi$ résulte de celle de la fonction f.

On calcule la formule

$$\sum_{x \in \omega_h} P_\xi(x)\,\xi(x)\,h = \sum_{x \in \omega_h} \{-\xi_{\bar{x}\dot{x}}(x) + f_{\bar{x}}(x, \ \xi_{\bar{x}}(x) + F_{\bar{x}}(x), \ \xi_{\dot{x}}(x) + F_{\dot{x}}(x))\}\,\xi(x)\,h.$$

On récrit la somme du second membre:

$$\begin{aligned} \sum_{x \in \omega_h} P_\xi(x)\,\xi(x)\,h = \sum_{x \in \omega_h} \{&-\xi_{\bar{x}\dot{x}}(x) + f_{\bar{x}}(x, \ \xi_{\bar{x}}(x) + F_{\bar{x}}(x), \ \xi_{\dot{x}}(x) + F_{\dot{x}}(x)) - f_{\bar{x}}(x, F_{\bar{x}}(x), F_{\dot{x}}(x))\}\,\xi(x)\,h + \\ &+ \sum_{x \in \omega_h} f_{\bar{x}}(x, F_{\bar{x}}(x), F_{\dot{x}}(x))\,\xi(x)\,h. \end{aligned}$$

Étant donné $\xi(0) = \xi(1) = 0$, la condition de monotonie forte et l'inégalité de Cauchy-Bouniakovski conduisent à

$$\sum_{x \in \omega_h} P_\xi(x)\, \xi(x)\, h \geqslant c_1 \sum_{x \in \bar{\omega}_h} \xi_{\bar{x}}^2(x)\, h - c_2 \left(\sum_{x \in \omega_h} \xi^2(x)\, h \right)^{1/2},$$

où

$$c_2 = \max_{x \in [0,1]} \left| f(x, F(x), F'(x)) \right| .$$

On s'est servi des propriétés suivantes de $F(x)$ linéaire :

$$F_{\bar{x}}(x) = F(x) \quad \text{et} \quad F_{\bar{x}}(x) = F'(x).$$

L'hypothèse $\xi(0) = \xi(1) = 0$ entraîne

$$\sum_{x \in \bar{\omega}_h} \xi_{\bar{x}}^2(x)\, h \geqslant \sum_{x \in \omega_h} \xi^2(x)\, h.$$

Donc

$$\sum_{x \in \omega_h} P_\xi(x)\, \xi(x)\, h \geqslant 8\, c_1 \sum_{x \in \omega_h} \xi^2(x)\, h - c_2 \left(\sum_{x \in \omega_h} \xi^2(x)\, h \right)^{1/2}.$$

On peut dire maintenant qu'avec σ choisi à partir de la condition $8\, c_1 \sigma^2 - c_2 \sigma \geqslant 0$,

$$\sum_{x \in \omega_h} P_\xi(x)\, \xi(x) \geqslant 0$$

si

$$\left(\sum_{x \in \omega_h} \xi^2(x) \right)^{1/2} = \rho = h^{-1/2} \sigma.$$

Conformément au lemme 6.1, le problème (6.9) possède donc nécessairement une solution telle que

$$\left(\sum_{x \in \omega_h} \xi^2(x)\, h \right)^{1/2} \leqslant \sigma.$$

THÉORÈME 6.2. *Si l'on est dans la condition* (6.3) *sous laquelle le problème* (6.1), (6.2) *possède une solution unique dans la classe* $C^{r+2}[0, 1]$, *la solution du problème approché* (6.6), (6.7) *admet le développement*

$$u^h(x) = u(x) + \sum_{j=1}^{l} h^{2j} v_j(x) + h^r \eta^h(x), \qquad x \in \bar{\omega}_h. \quad (6.10)$$

Ici $l = [(r-1)/2]$, *les fonctions* v_j *sont dans* $C^{r+2-2j}[0, 1]$ *et ne dépendent pas de* h, *la fonction discrète* η^h *est uniformément bornée*:

$$|\eta^h(x)| \leqslant c_3 \qquad \forall\, x \in \bar{\omega}_h, \tag{6.11}$$

avec la constante c_3 *indépendante de* h.

DÉMONSTRATION. On pose $v_0 = u$ sur $[0, 1]$ et on cherche l fonctions v_j à partir des problèmes différentiels

$$-v_j'' + \frac{\partial f}{\partial v}(x, u(x), u'(x))\, v_j' + \frac{\partial f}{\partial u}(x, u(x), u'(x))\, v_j =$$
$$= -\frac{1}{4^j (2j)!}\frac{d^{2j}}{dx^{2j}} f(x, u, u') + \sum_{k=1}^{j} \frac{2 v_{j-k}^{(2k+2)}}{(2k+2)!} - \frac{\partial f}{\partial v}(x, u, u') \times$$
$$\times \sum_{k=1}^{j} \frac{v_{j-k}^{(2k+1)}}{4^k (2k+1)!} - \frac{\partial f}{\partial u}(x, u, u') \sum_{k=1}^{j} \frac{v_{j-k}^{(2k)}}{4^k (2k)!} -$$
$$- \sum_{\substack{2 \leqslant s+m+p \leqslant j \\ s+m \geqslant 1}} \frac{1}{4^p (2p)!} \frac{d^{2p}}{dx^{2p}} \Bigg(\frac{1}{s!\, m!} \frac{\partial^{s+m} f}{\partial u^s \partial v^m}(x, u, u') \times$$
$$\times \sum_{\substack{t_1 + \ldots + t_{s+m} = j-p \\ t_i \geqslant 1}} \prod_{i=1}^{s} \left(\sum_{k=0}^{t_i} \frac{v_{t_i-k}^{(2k)}}{4^k (2k)!} \right) \prod_{i=s+1}^{s+m} \left(\sum_{k=0}^{t_i} \frac{v_{t_i-k}^{(2k+1)}}{4^k (2k+1)!} \right) \Bigg), \tag{6.12}$$

$$v_j(0) = 0, \qquad v_j(1) = 0, \qquad j = 1, 2, \ldots, l. \tag{6.13}$$

On commence par v_1. L'équation (6.12) a alors la forme la plus simple, à savoir

$$-v_1'' + \frac{\partial f}{\partial v}(x, u, u')\, v_1' + \frac{\partial f}{\partial u}(x, u, u')\, v_1 = -\frac{\partial f}{\partial v}(x, u, u') \times$$
$$\times \frac{v_0'''}{24} - \frac{\partial f}{\partial u}(x, u, u') \frac{v_0''}{8} - \frac{1}{8}\frac{d^2}{dx^2} f(x, u, u') + \frac{v_0^{(4)}}{12}.$$

Le second membre comprend seulement v_0 et admet des dérivées continues jusqu'à l'ordre $r-2$. La dernière propriété des coefficients et le caractère linéaire du problème entraînent que la solution remplissant les conditions aux limites (6.13) existe, est unique et appartient à $C^r[0, 1]$. On suppose maintenant que $v_0, \ldots, v_{j-1}$ sont déjà définies (on observe que $v_k \in C^{r+2-2k}[0, 1]$). Les équations (6.12) montrent que le second membre de l'équation en v_j renferme

v_k d'indice $j-1$ au plus et qu'il est $r-2j$ fois continûment dérivable sur le segment [0, 1]. On établit la dernière propriété en calculant les indices de régularité des fonctions v_k et leurs ordres de dérivation le plus élevés. Le problème linéaire ainsi construit a pour coefficients des fonctions de classe $C^{r-1}[0, 1]$, si bien qu'il existe une seule solution qui vérifie, en plus de (6.12), les conditions aux limites (6.13). En outre, $v_j \in C^{r+2-2j}[0, 1]$. Ainsi, on a obtenu toutes les $l+1$ fonctions v_j telles que

$$v_j \in C^{r-2j+2}[0, 1], \qquad j = 0, 1, \ldots, l.$$

On définit par v_j les fonctions

$$w(x) = \sum_{j=0}^{l} h^{2j} v_j(x), \qquad x \in \bar{\omega}_h, \tag{6.14}$$

et on porte $w(x)$ dans l'opérateur aux différences du problème (6.6):

$$-\sum_{j=0}^{l} h^{2j}(v_j)_{\bar{x}x} + f_{\bar{x}}\left(x, \sum_{j=0}^{l} h^{2j}(v_j)_{\bar{x}}, \sum_{j=0}^{l} h^{2j}(v_j)_{x}\right) =$$
$$= -w_{\bar{x}x} + f_{\bar{x}}(x, w_{\bar{x}}, w_{x}). \tag{6.15}$$

On transforme les termes à l'aide des développements du lemme 1.1, § 7.1:

$$-\sum_{j=0}^{l} h^{2j}(v_j)_{\bar{x}x} = -2\sum_{j=0}^{l} h^{2j}\sum_{k=0}^{l-j} h^{2k}\frac{v_j^{(2k+2)}}{(2k+2)!} + h^r\theta_1 =$$
$$= -2\sum_{j=0}^{l} h^{2j}\sum_{k=0}^{j}\frac{v_{j-k}^{(2k+2)}}{(2k+2)!} + h^r\theta_1 \quad \text{sur} \quad \omega_h.$$

$$\sum_{j=0}^{l} h^{2j}(v_j)_{\bar{x}} = \sum_{j=0}^{l} h^{2j}\sum_{k=0}^{l-j} h^{2k}\frac{v_j^{(2k)}}{2^{2k}(2k)!} + h^r\theta_2 =$$
$$= \sum_{j=0}^{l} h^{2j}\sum_{k=0}^{j}\frac{v_{j-k}^{(2k)}}{2^{2k}(2k)!} + h^r\theta_2 \quad \text{sur} \quad \bar{\omega}_h. \tag{6.16}$$

$$\sum_{j=0}^{l} h^{2j}(v_j)_{\bar{x}} - \sum_{j=0}^{l} h^{2j}\sum_{k=0}^{l-j} h^{2k}\frac{v_j^{(2k+1)}}{2^{2k}(2k+1)!} + h^r\theta_3 =$$
$$= \sum_{j=0}^{l} h^{2j}\sum_{k=0}^{j}\frac{v_{j-k}^{(2k+1)}}{2^{2k}(2k+1)!} + h^r\theta_3 \quad \text{sur} \quad \bar{\omega}_h,$$

et

$$|\theta_i| \leqslant c_4 \quad \text{pour} \quad i = 1,\ 2,\ 3.$$

On développe $f(x,\ w_{\bar{x}},\ w_{\mathring{x}})$ en la formule de Taylor:

$$f\left(x,\ \sum_{j=0}^{l} h^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k)}}{4^k (2k)!} + h^r\theta_2,\ \sum_{j=0}^{l} h^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k+1)}}{4^k (2k+1)!} + \right.$$

$$\left. + h^r\theta_3\right) = f(x,\ u,\ u') + \sum_{1 \leqslant s+m \leqslant l} \frac{1}{s!\,m!} \frac{\partial^{s+m}}{\partial u^s \partial v^m} f(x,\ u,\ u') \times$$

$$\times \left(\sum_{j=1}^{l} h^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k)}}{4^k (2k)!} + h^r\theta_2\right)^s \left(\sum_{j=1}^{l} h^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k+1)}}{4^k (2k+1)!} + h^r\theta_3\right)^m +$$

$$+ h^{2l+2}\theta_4 \text{ sur } \omega_h. \qquad (6.17)$$

où

$$\theta_4 = \sum_{s+m=l+1} \frac{1}{s!\,m!} \frac{\partial^{l+1} f}{\partial u^s \partial v^m} (x,\ \xi_s,\ \eta_m) \times$$

$$\times \left(\sum_{j=1}^{l} h^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k)}}{4^k (2k)!} + h^r\theta_2\right)^s \left(\sum_{j=1}^{l} h^{2j} \sum_{k=0}^{j} \frac{v_{j-k}^{(2k+1)}}{4^k (2k+1)!} + h^r\theta_3\right)^m.$$

La quantité θ_4 est uniformément bornée:

$$|\theta_4| \leqslant c_5$$

parce que les fonctions v_j sont bornées en chaque point $x \in \omega_h$ (par suite de leur continuité sur [0, 1]), ainsi que leurs dérivées correspondantes de (6.17).

On fait le produit de toutes les parenthèses de (6.17) en tant que polynômes en h:

$$f(x,\ w_{\bar{x}},\ w_{\mathring{x}}) = f(x,\ u,\ u') +$$

$$+ \sum_{1 \leqslant s+m \leqslant l} \frac{1}{s!\,m!} \frac{\partial^{s+m} f}{\partial u^s \partial v^m} (x,\ u,\ u') \sum_{j=s+m}^{l} h^{2j} \times$$

$$\times \left(\sum_{\substack{t_1+\ldots+t_{s+m}=j \\ t_i \geqslant 1}} \prod_{i=1}^{s} \left(\sum_{k=0}^{t_i} \frac{v_{t_i-k}^{(2k)}}{4^k (2k)!}\right) \prod_{i=s+1}^{s+m} \left(\sum_{k=0}^{t_i} \frac{v_{t_i-k}^{(2k+1)}}{4^k (2k+1)!}\right)\right) + h^r\theta_5.$$

et on change l'ordre de sommation, il vient

$$f(x, w_{\bar{x}}, w_{\hat{x}}) =$$

$$= f(x, u, u') + \sum_{j=1}^{l} h^{2j} \sum_{1 \leqslant s+m \leqslant j} \frac{1}{s!\,m!} \frac{\partial^{s+m} f}{\partial u^s \partial v^m}(x, u, u') \times$$

$$\times \left(\sum_{\substack{l_1+\dots+l_{s+m}=j \\ l_i \geqslant 1}} \prod_{i=1}^{s} \left(\sum_{k=0}^{l_i} \frac{v_{l_i-k}^{(2k)}}{4^k (2k)!} \right) \prod_{i=s+1}^{s+m} \left(\sum_{k=0}^{l_i} \frac{v_{l_i-k}^{(2k+1)}}{4^k (2k+1)!} \right) \right) +$$

$$+ h^r \theta_5 \quad \text{sur} \quad \bar{\omega}_h.$$

On transforme la demi-somme des valeurs de f par la formule (1.2) du § 7.1 :

$$f_{\hat{x}}(x, w_{\bar{x}}, w_{\hat{x}}) = f(x, u, u') + \sum_{j=1}^{l} h^{2j} \left\{ \frac{1}{4^j (2j)!} \frac{d^{2j}}{dx^{2j}} f(x, u, u') + \right.$$

$$+ \sum_{p=0}^{l-j} h^{2p} \frac{1}{4^p (2p)!} \frac{d^{2p}}{dx^{2p}} \left(\sum_{1 \leqslant s+m \leqslant j} \frac{1}{s!\,m!} \frac{\partial^{s+m} f}{\partial u^s \partial v^m}(x, u, u') \times \right.$$

$$\left. \left. \times \sum_{\substack{l_1+\dots+l_{s+m}=j \\ l_i \geqslant 1}} \prod_{i=1}^{s} \left(\sum_{k=0}^{l_i} \frac{v_{l_i-k}^{(2k)}}{4^k (2k)!} \right) \sum_{i=s+1}^{s+m} \left(\sum_{k=0}^{l_i} \frac{v_{l_i-k}^{(2k+1)}}{4^k (2k+1)!} \right) \right) \right\} + h^r \theta_6.$$

L'égalité est juste sur le réseau ω_h tout entier. On range les termes dans l'ordre de croissance des puissances de h et on porte dans (6.15) compte tenu de (6.16). On a

$$- w_{\bar{x}\hat{x}} + f_{\hat{x}}(x, w_{\bar{x}}, w_{\hat{x}}) = - v_0'' + f(x, u, u') +$$

$$+ \sum_{j=1}^{l} h^{2j} \left\{ \frac{1}{4^j (2j)!} \frac{d^{2j}}{dx^{2j}} f(x, u, u') - 2 \sum_{k=0}^{j} \frac{v_{j-k}^{(2k+2)}}{(2k+2)!} + \right.$$

$$+ \sum_{p=0}^{j} \frac{1}{4^p (2p)!} \frac{d^{2p}}{dx^{2p}} \left(\sum_{1 \leqslant s+m \leqslant j-p} \frac{1}{s!\,m!} \frac{\partial^{s+m} f}{\partial u^s \partial v^m}(x, u, u') \right) \times$$

$$\left. \times \sum_{\substack{l_1+\dots+l_{s+m}=j-p \\ l_i \geqslant 1}} \prod_{i=1}^{s} \left(\sum_{k=0}^{l_i} \frac{v_{l_i-k}^{(2k)}}{4^k (2k)!} \right) \prod_{i=s+1}^{m+s} \left(\sum_{k=0}^{l_i} \frac{v_{l_i-k}^{(2k+1)}}{4^k (2k+1)!} \right) \right\} + h^r \theta_7,$$

où

$$|\theta_7| \leqslant c_6 \qquad \forall x \in \omega_h.$$

On note qu'étant donné $v_0 = u$, le terme $-v_0'' + f(x, u, u')$ s'annule. En vertu de (6.12), tous les termes en h^{2j}, $j = 1, \ldots, l$, se réduisent mutuellement. Aussi

$$-w_{\bar{x}\hat{x}} + f_{\bar{x}}(x, w_{\bar{x}}, w_{\hat{x}}) = h^r \theta_8 \quad \text{sur} \quad \omega_h. \tag{6.18}$$

Ici $\theta_8(x)$ désigne une fonction discrète uniformémemt bornée en valeur absolue par une constante indépendante de h:

$$|\theta_8(x)| \leqslant c_7 \qquad \forall x \in \omega_h. \tag{6.19}$$

Il y a lieu d'observer que les conditions aux limites homogènes (6.13) et la définition (6.14) de w entraînent la validité des égalités $w(0) = u^h(0) = u_0$ et $w(1) = u^h(1) = u_1$, si bien qu'on utilise la propriété de monotonie forte (6.8), où l'on pose $u(x) = u^h(x)$ et $v(x) = w(x)$. On a en raison de (6.6) et (6.18):

$$-h^r \sum_{x \in \omega_h} \theta_8(x)\left(u^h(x) - w(x)\right) h \geqslant c_1 \sum_{x \in \bar{\omega}_h} \left(u^h_{\bar{x}}(x) - w_{\bar{x}}(x)\right)^2 h. \tag{6.20}$$

Comme $u^h(0) - w(0) = 0$ et $u^h(1) - w(1) = 0$, on utilise l'analogue aux différences de l'estimation de l'immersion de $\mathring{W}_2^1(0, 1)$ dans $C[0, 1]$ (voir [41]):

$$\max_{x \in \bar{\omega}_h} |u^h(x) - w(x)| \leqslant \frac{1}{2}\left(\sum_{x \in \bar{\omega}_h} \left(u^h_{\bar{x}}(x) - w_{\bar{x}}(x)\right)^2 h\right)^{1/2}. \tag{6.21}$$

Il résulte de plus de (6.19)

$$\sum_{x \in \omega_h} \theta_8(x)\,|u^h(x) - w(x)|\, h \leqslant c_7 \max_{x \in \bar{\omega}_h} |u^h(x) - w(x)|. \tag{6.22}$$

On réunit (6.20) et (6.21):

$$4c_1\left(\max_{x \in \bar{\omega}_h} |u^h(x) - w(x)|\right)^2 \leqslant c_7 h^r \max_{x \in \bar{\omega}_h} |u^h(x) - w(x)|.$$

La division par $\max |u^h(x) - w(x)|$ donne

$$\max_{x \in \bar{\omega}_h} |u^h(x) - w(x)| \leqslant \frac{c_7}{4c_1} h^r.$$

Ainsi, avec la notation

$$\eta^h(x) = \left(u^h(x) - w(x)\right) h^{-r}, \qquad x \in \bar{\omega}_h,$$

on a pour cette fonction discrète

$$\max_{x \in \bar{\omega}_h} |\eta^h(x)| \leqslant \frac{c_7}{4c_1},$$

inégalité qui coïncide avec la majoration (6.11) à condition de poser

$$c_3 = \frac{c_7}{4\,c_1}.$$

Le théorème 6.2 se trouve démontré.

Voici un exemple qui utilise le développement du théorème 6.2. Soit $l = [(r - 1)/2]$. On construit pour $l + 1$ entiers $N_k = kN$, $k = 1, \ldots, l + 1$, les réseaux $\bar{\omega}_{h_k}$ de pas $h_k = h/k$, où $h = 1/N$, et on cherche les solutions des problèmes aux différences (6.6), (6.7). Comme il y a unicité, des algorithmes de recherche différents fournissent une même solution. On trouve dans [119] de nombreux procédés itératifs pour les systèmes non linéaires de la forme (6.6), (6.7).

Toutes les solutions u^{h_k} sont définies sur le réseau $\bar{\omega}_h$ de pas h. On additionne leurs valeurs nodales avec les poids

$$\gamma_k = 2\,\frac{(-1)^{l-k+1}\,k^{2l+2}}{(l-k+1)!\,(l+k+1)!}. \qquad (6.23)$$

La solution corrigée s'écrit

$$U = \sum_{k=1}^{l+1} \gamma_k u^{h_k} \quad \text{sur} \quad \bar{\omega}_h. \qquad (6.24)$$

THÉORÈME 6.3. *On est dans les hypothèses du théorème* 6.2. *La solution corrigée* (6.24) *admet l'estimation*

$$\max_{x \in \bar{\omega}_h} |U(x) - u(x)| \leqslant c_8 h^r,$$

la constante c_8 étant indépendante de h et k.

La démonstration est calquée sur celle du théorème 4.6, et

$$c_8 = 2\,c_3 \sum_{k=1}^{l+1} \frac{k^{2l+2-r}}{(l-k+1)!\,(l+k+1)!}.$$

Le gain en précision est illustré par l'exemple numérique suivant. Soit le problème

$$\begin{gathered} u'' = \cos(u' + 4u) + 16\,e^{-4x} - \cos(4x + 1), \qquad x \in (0, 1), \\ u(0) = 1, \qquad u(1) = 1 + e^{-4}. \end{gathered} \qquad (6.25)$$

Il a pour solution exacte (et unique) la fonction

$$u(x) = e^{-4x} + x.$$

On a construit pour un jeu de N_i entiers les réseaux ω_{h_i} et calculé les solutions u^{h_i} des problèmes aux différences (6.6), (6.7). On a cherché u^{h_i} par itérations, si bien qu'on a abouti à v^{h_i} distincte de u^{h_i}. On a utilisé, pour éliminer l'erreur $v^{h_i} - u^{h_i}$, un

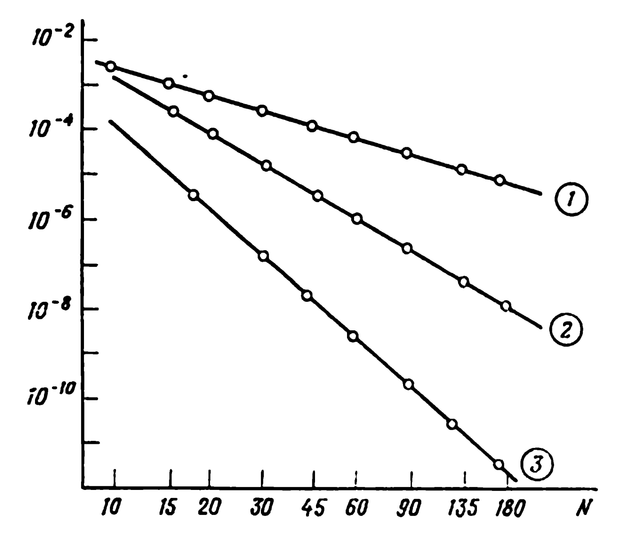


Fig. 3.4. Erreurs maxima sur les solutions approchées du problème (6.25)

1 — erreur sur la solution aux différences du problème (6.6), (6.7); *2* — erreur sur la solution extrapolée sur deux réseaux pour le rapport de pas 1:2; *3* — erreur sur la solution extrapolée sur trois réseaux pour le rapport de pas 1:2:3.

procédé itératif qui garantit une précision supérieure à 10^{-12}, i.e.

$$\left| v^{h_i}(x) - u^{h_i}(x) \right| \leqslant 10^{-12} \qquad \forall\, x \in \omega_{h_i}.$$

A cet effet, on a calculé l'erreur maximum

$$\xi(N_i) = \max_{x \in \omega_{h_i}} \left| u^{h_i}(x) - u(x) \right|,$$

dont le graphe est représenté sur la fig. 3.4. On a construit enfin les solutions améliorées (6.24) en extrapolant sur deux ou trois solutions pour le rapport de pas 1 : 2 et 1 : 2 : 3 respectivement. La fig. 3.4 donne la dépendance des erreurs sur les solutions améliorées (6.24) par rapport au nombre total de points d'extrapolation.

CHAPITRE 4

ÉQUATIONS DU TYPE ELLIPTIQUE

Les équations du type elliptique avec les conditions aux limites correspondantes forment la plus vaste classe de problèmes de la physique mathématique. Elles interviennent dans beaucoup d'applications ayant un intérêt pratique direct et dans la réduction temporelle des problèmes paraboliques et hyperboliques. On conçoit que les problèmes liés aux opérateurs elliptiques ont à nos yeux une importance particulière. Il n'est pas dans nos intentions d'insister sur les résultats connus ayant trait à la position des problèmes elliptiques et à la façon dont les propriétés de leurs solutions dépendent de celles des entrées. Nous renvoyons pour ces résultats aux ouvrages cités dans la Bibliographie *in fine*. Nous voulons donner le strict nécessaire sur le plan théorique tout en centrant notre attention sur divers procédés de raffinement des solutions aux différences. En plus des problèmes linéaires, nous étudierons un problème non linéaire avec les conditions de possibilité. On a décidé d'utiliser des cas relativement simples pour mettre en lumière les moyens de raffinement et familiariser le lecteur avec les procédés de justification des algorithmes (et on a choisi en conséquence le problème non linéaire et le problème de diffraction). Ces idées se généralisent visiblement à une classe plus vaste de problèmes, mais cela nous entraînerait hors du cadre que nous nous sommes fixé.

4.1. Positions des problèmes différentiels étudiés

Soit $\mathbf{R}^2$ l'espace euclidien de dimension 2 de point générique $\mathbf{x} = (x_1, x_2) = (x, y)$ muni de la distance

$$|\mathbf{x} - \mathbf{x}'| = ((x_1 - x_1')^2 + (x_2 - x_2')^2)^{1/2},$$

Ω un domaine borné connexe strictement lipschitzien (voir [26]) de frontière Γ. On étudiera tout le long du chapitre l'équation

$$Lu \equiv -\frac{\partial}{\partial x_1} p \frac{\partial u}{\partial x_1} - \frac{\partial}{\partial x_2} p \frac{\partial u}{\partial x_2} + qu = f \quad \text{sur} \quad \Omega, \tag{1.1}$$

avec les coefficients vérifiant les conditions

$$p \geq c_1 > 0, \qquad q \geq 0 \quad \text{sur} \quad \Omega. \tag{1.2}$$

Quant aux conditions aux limites, nous considérerons le problème de Dirichlet

$$u(\mathbf{x}) = \varphi(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \Gamma. \tag{1.3}$$

Soit $\alpha \in (0, 1)$. On introduit la notation

$$< u >_{\Omega}^{\alpha} = \sup_{\mathbf{x}, \mathbf{x}' \in \Omega} \frac{|u(\mathbf{x}) - u(\mathbf{x}')|}{|\mathbf{x} - \mathbf{x}'|^{\alpha}}.$$

Les classes de régularité ci-dessous sont définies conformément à [26].

$C^{l+\alpha}(\bar{\Omega})$, espace de Banach des fonctions continues dans $\bar{\Omega}$ qui possèdent dans Ω des dérivées continues jusqu'à l'ordre l inclus telles que la quantité

$$\| u \|_{C^{l+\alpha}\bar{\Omega}} = \sum_{0 \leq k_1 + k_2 \leq l} \sup_{\Omega} \left| \frac{\partial^{k_1+k_2} u}{\partial x_1^{k_1} \partial x_2^{k_2}} \right| + \sum_{k_1+k_2=l} \left\langle \frac{\partial^{k_1+k_2} u}{\partial x_1^{k_1} \partial x_2^{k_2}} \right\rangle_{\Omega}^{\alpha},$$

$k_i \geq 0$ étant des entiers, soit finie;

$C^{l+\alpha}(\Omega)$, ensemble des fonctions de $C^{l+\alpha}(\bar{\Omega}')$ pour tout $\Omega' \subset \Omega$ strictement intérieur (i.e. tel que la distance entre Ω' et Γ soit positive);

$L_2(\Omega)$, espace de Banach des fonctions mesurables de carré sommable dans Ω au sens de Lebesgue. On note $\| u \|_{L_2(\Omega)} = \left(\int_{\Omega} u^2 \, d\mathbf{x} \right)^{1/2}$;

$W_2^l(\Omega)$, espace de Banach de toutes les fonctions $u \in L_2(\Omega)$ ayant des dérivées généralisées $\partial^{k_1+k_2} u / \partial x_1^{k_1} \partial x_2^{k_2}$, $0 \leq k_1 + k_2 \leq l$, de carré sommable sur Ω. Les dérivées généralisées sont comprises au sens classique (par exemple, comme dans [34], [135]). La norme est définie par l'égalité $\| u \|_{W_2^l(\Omega)} = \int_{\Omega} \left(\sum_{0 \leq k_1 + k_2 \leq l} \frac{\partial^{k_1+k_2} u}{\partial x_1^{k_1} \partial x_2^{k_2}} d\mathbf{x} \right)^{1/2}$;

$\overset{\circ}{W}{}_2^l(\Omega)$, sous-espace de $W_2^l(\Omega)$ qui est la fermeture pour la norme $\| \cdot \|_{W_2^l(\Omega)}$ de toutes les fonctions indéfiniment dérivables à support dans Ω.

Il y a des fois intérêt à munir $\overset{\circ}{W}{}_2^1(\Omega)$ d'une norme définie comme suit:

$$| u | = \left(\int_{\Omega} \left\{ \left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right\} d\mathbf{x} \right)^{1/2}.$$

Les normes $\| u \|_{\overset{\circ}{W}{}_2^1(\Omega)}$ et $| u |$ sont équivalentes (voir [34]).

Une courbe K est dite *de classe* $C^{k+\alpha}$ (resp. C^k) si l'on trouve un nombre $a_0 > 0$ tel qu'on introduise au voisinage de chaque point $\mathbf{x}_0 \in K$ les coordonnées cartésiennes (y_1, y_2) de centre $\mathbf{x}_0$ dans lesquelles

les points de la frontière sont décrits par l'équation $y_2 = g(y_1)$, la fonction $g(y_1)$ étant de classe $C^{k+\alpha}[a, b]$ (resp. $C^k [a, b]$), avec $[a, b]$ obtenu par projection de l'ensemble $\{\mathbf{x} \in K;\ |\mathbf{x} - \mathbf{x}_0| \leqslant a_0\}$ sur la droite $y_2 = 0$.

On donne (sans démonstration) le résultat de possibilité du problème (1.1) à (1.3).

THÉORÈME 1.1 (voir [26]). *On suppose que les coefficients du problème* (1.1) *à* (1.3)

$$p \in C^{l+1+\alpha}(\bar{\Omega}), \qquad q,\ f \in C^{l+\alpha}(\bar{\Omega}), \tag{1.4}$$

avec $l \geqslant 0$ un entier et $\alpha \in (0, 1)$. Il existe une solution unique $u \in C^{l+2+\alpha}(\Omega)$.

Si l'on ajoute aux hypothèses du théorème la régularité de la frontière et des valeurs limites, on garantit la régularité « jusqu'à la frontière » de la solution.

THÉORÈME 1.2 (voir [26]). *On suppose que les coefficients du problème* (1.1) *à* (1.3) *satisfont aux conditions* (1.4) *et que*

$$\Gamma \in C^{l+2+\alpha}, \qquad \varphi \in C^{l+2+\alpha}(\Gamma).$$

On a $u \in C^{l+2+\alpha}(\bar{\Omega})$.

Le problème (1.1) à (1.3) a été formulé en termes d'opérateurs. On en donne un énoncé (intégral) faible. On appelle *solution généralisée* (voir [26]) *du problème* (1.1) *avec la condition aux limites* (1.3) *homogène* (i.e. $\varphi \equiv 0$) une fonction $u \in \mathring{W}_2^1(\Omega)$ vérifiant l'identité intégrale

$$\int_\Omega \left(p \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + p \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2} + quv\right) d\mathbf{x} = \int_\Omega fv\, d\mathbf{x} \qquad \forall v \in \mathring{W}_2^1(\Omega). \tag{1.5}$$

Si le coefficient p admet des dérivées généralisées du premier ordre et $u \in W_2^2(\Omega)$, l'identité (1.5) découle de l'équation (1.1), et vice versa. Voici un critère de possibilité du problème (1.5).

THÉORÈME 1.3 (voir [26]). *On suppose que les coefficients p, q de l'équation* (1.5) *sont des fonctions mesurables bornées sur $\bar{\Omega}$ telles qu'on ait* (1.2). *Il existe pour toute fonction f de norme $\|f\|_{L_2(\Omega)}$ bornée une solution généralisée unique du problème* (1.5).

On démontre dans certains cas une régularité plus grande.

THÉORÈME 1.4. *On suppose que les coefficients du problème* (1.5) *vérifient les conditions du théorème* 1.3, *que*

$$\left|\frac{\partial p}{\partial x_i}\right| \leqslant c_2 < \infty, \qquad i = 1, 2,$$

et que, ou bien Ω est l'un des domaines suivants : un disque, une couronne circulaire, un rectangle, un triangle, ou bien il le devient par une transformation régulière de classe C^2 $(\bar{\Omega})$. Le problème admet une solution unique de classe W_2^2 (Ω).

DÉMONSTRATION. S'agissant du disque, de la couronne circulaire, du rectangle et des domaines qui les deviennent par des transformations, l'affirmation est donnée par [26]. Soit Ω un triangle ouvert. On se place dans le cas

$$-\Delta v = f \quad \text{dans} \quad \Omega, \qquad v = 0 \quad \text{sur} \quad \Gamma. \tag{1.6}$$

On complète Ω par son symétrique par rapport au plus grand côté et on adjoint les points de ce côté, il vient un quadrilatère convexe $\tilde{\Omega}$ de frontière $\tilde{\Gamma}$. On prolonge la fonction f de façon antisymétrique au triangle symétrique de Ω et on la prolonge par zéro à l'axe de symétrie, il vient le problème

$$-\Delta \tilde{v} = f \quad \text{dans} \quad \tilde{\Omega}, \qquad \tilde{v} = 0 \quad \text{sur} \quad \tilde{\Gamma}. \tag{1.7}$$

Le quadrilatère convexe $\tilde{\Omega}$ se ramène par une transformation régulière bilinéaire à un rectangle, et la fonction prolongée f est dans L_2 $(\tilde{\Omega})$, si bien que le problème (1.7) admet une solution unique $\tilde{v} \in W_2^2$ $(\tilde{\Omega})$. On démontre que $\tilde{v}$ vérifie sur Ω le problème (1.6). En effet, $\tilde{v}$ vérifie presque partout dans Ω l'équation $-\Delta\tilde{v} = f$ et $\tilde{v} \in W_2^2\,(\Omega)$. De plus, les valeurs de $\tilde{v}$ sont antisymétriques par rapport à l'axe de symétrie introduit, ce qui fait que $\tilde{v}$ s'annule sur le plus grand côté du triangle Ω. Cette fonction est également nulle sur les deux autres côtés de Ω (par suite de la condition aux limites homogène (1.7)). Ainsi, la solution trouvée de (1.6) est bien unique.

Les raisonnements suivants basés sur la transformation de variables répètent mot pour mot ceux de [26].

4.2. Méthodes par différences finies pour le problème de Dirichlet dans un domaine de frontière régulière

Les méthodes par différences finies de ce paragraphe fonctionnent pour une classe assez vaste d'équations quasi linéaires et d'équations à plusieurs variables encore que la démonstration des grands théorèmes et l'obtention des autres résultats deviennent trop laborieuses. On se limitera donc à des problèmes aux limites simples pour des équations du type elliptique, ce qui permettra de s'appesantir sur les idées des procedés de raffinement des solutions approchées.

Ce paragraphe est consacré au problème de Dirichlet pour l'équation de Poisson en dimension deux

$$-\Delta u = f \quad \text{dans} \quad \Omega, \tag{2.1}$$

$$u = \varphi \quad \text{sur} \quad \Gamma. \tag{2.2}$$

On suppose qu'on est dans les hypothèses du théorème 1.2 pour $l \geqslant 2$ entier et un certain α de l'intervalle $(0, 1)$.

4.2.1. Approximation de la condition aux limites

L'idée de base de ce numéro est justifiée dans [9].

On discrétise le problème posé en supposant que le domaine Ω est inclus dans le carré $\{-b < x < b, \ -b < y < b\}$ qu'on recouvre d'un réseau carré de pas $h = b/N$ formé par les lignes $x_i = ih$, $y_j = jh$, avec $i, j = -N, \ldots, N$. Les points d'intersection de ces droites s'appellent les *nœuds*. Un nœud $\mathbf{x} = (x_i, y_j)$ est dit *intérieur* si $\mathbf{x} \in \Omega$. On désigne par Ω_h l'ensemble des nœuds intérieurs. Chaque fois qu'une ligne du réseau traverse Ω, elle traverse Γ, et son intersection avec le domaine contient plusieurs intervalles (car $\Gamma \in C^{4+\alpha}$). Selon que la ligne concernée est parallèle à l'axe Ox ou à l'axe Oy, les extrémités de ces intervalles constituent des *nœuds frontières* dans la direction x ou y. On note que le point A de la fig. 4.1 est un nœud frontière dans la direction x bien qu'il soit sur une ligne parallèle à l'axe Oy. On appelle $\Gamma_{h,x}$ et $\Gamma_{h,y}$ les ensembles des nœuds frontières dans la direction x et dans la direction y respectivement. On notera Γ_h la réunion de ces ensembles, i.e. l'ensemble des nœuds frontières. Un nœud intérieur est dit *régulier* si, chaque fois que $\bar{\Omega}$ le contient, elle contient quatre segments formés des lignes du réseau, qui le réunissent à quatre nœuds le plus proches de Ω_h. On note Ω_h^r l'ensemble des nœuds réguliers. Les autres nœuds intérieurs sont dits *irréguliers*. Ils forment l'ensemble Ω_h^{ir}.

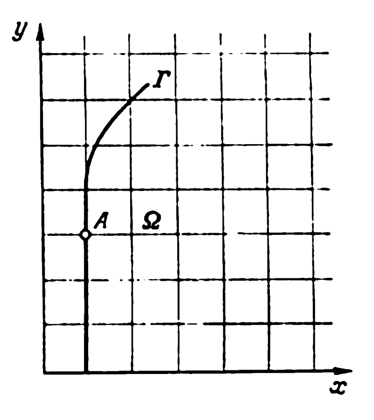


Fig. 4.1. Nœud frontière dans la direction x

On fait correspondre à chaque nœud régulier une équation aux différences usuelle à cinq points

$$-u^h_{x\bar{x}} - u^h_{y\bar{y}} = f \quad \text{sur} \quad \Omega_h^r \tag{2.3}$$

qui donne pour $u \in C^{l+\alpha+2}(\bar{\Omega})$ l'erreur d'approximation

$$-u_{\bar{x}\mathring{x}} - u_{\bar{y}\mathring{y}} = -\Delta u - \sum_{k=1}^{\left[\frac{l}{2}\right]} h^{2k} \frac{2}{(2k+2)!}\left(\frac{\partial^{2k+2} u}{\partial x^{2k+2}} + \frac{\partial^{2k+2} u}{\partial y^{2k+2}}\right) + \frac{2h^{l+\alpha}}{(2l+2)!}\theta_1, \quad (2.4)$$

où

$$|\theta_1(\mathbf{x})| \leqslant \|u\|_{C^{l+2+\alpha}(\bar{\Omega})} \qquad \forall \mathbf{x} \in \Omega_h^r.$$

On note que les coefficients des puissances paires de h sont indépendants de h et constituent des fonctions régulières sur $\bar{\Omega}$. Les résultats du § 1.2 autorisent donc à croire que ces termes de l'erreur d'approximation interviennent de même façon dans l'erreur sur la solution. S'agissant des équations associées aux nœuds irréguliers, on se heurte toutefois à une difficulté. Le fait est que les équations usuelles à cinq points basées sur des nœuds irrégulièrement espacés ne donnent pas lieu à l'erreur d'approximation requise, si bien qu'on les remplace par des équations discrètes provenant de la formule d'interpolation de Lagrange.

Soit, par exemple, sur l'axe Ot le point $\delta\bar{h}$ à droite de l'origine et n points équidistants $-\bar{h}, -2\bar{h}, \ldots, -n\bar{h}$ à gauche. L'interpolation par Lagrange à partir des valeurs $\psi(t)$ en ces points fournit la formule

$$\psi(0) = \sum_{k=1}^{n}(-1)^{k-1}\frac{n!}{k!\,(n-k)!}\,\frac{\delta}{\delta+k}\,\psi(-k\bar{h}) + \\ + \prod_{k=1}^{n}\frac{k}{\delta+k}\,\psi(\delta\bar{h}) + R(0), \quad (2.5)$$

avec le reste

$$|R(0)| \leqslant \bar{h}^{n+1}\frac{\delta}{n+1}\max_{t\in[-n\bar{h},\,\delta\bar{h}]}\left|\frac{d^{n+1}\psi(t)}{dt^{n+1}}\right|. \quad (2.6)$$

On introduit les notations suivantes: l_δ^n est l'opérateur

$$l_\delta^n\,\psi(0) = \sum_{k=1}^{n}(-1)^{k-1}\frac{n!}{k!\,(n-k)!}\,\frac{\delta}{\delta+k}\,\psi(-k\bar{h}) \quad (2.7)$$

et λ_δ^n le nombre

$$\lambda_\delta^n = \prod_{k=1}^{n}\frac{k}{\delta+k}.$$

Puisque la quantité

$$B(\delta) = \sum_{k=1}^{n} \frac{n!}{k!(n-k!)} \frac{\delta}{\delta + k}$$

tend de façon monotone vers 0 avec $\delta > 0$, il existe δ_n tel que

$$B(\delta) < \frac{1}{2} \qquad \forall \delta \leqslant \delta_n. \tag{2.8}$$

On applique la formule d'interpolation obtenue à chaque $\mathbf{x} \in \Omega_h^{ir}$ en prenant l'axe Ot parallèle à l'un des axes de coordonnées et en choisissant comme origine le point $\mathbf{x}$. Le nœud $\mathbf{x}$ est bien irrégulier : parmi les quatre nœuds le plus proches de $\mathbf{x}$, il existe au moins un seul nœud frontière $\xi_\mathbf{x}$ dont la distance à $\mathbf{x}$ est inférieure à h. On dirige l'axe Ot de $\mathbf{x}$ vers $\xi_\mathbf{x}$ et on pose $\delta\bar{h}$ égale à la distance de $\mathbf{x}$ à $\xi_\mathbf{x}$. On prend le pas

$$\bar{h} = h\,([1/\delta_n] + 1). \tag{2.9}$$

L'équation correspondant au cas irrégulier s'écrit

$$u^h(\mathbf{x}) = I_\delta^n u^h(\mathbf{x}) + \lambda_\delta^n \varphi(\xi_\mathbf{x}), \; \mathbf{x} \in \Omega_h^{ir}, \tag{2.10}$$

chaque point $\mathbf{x}$ vérifiant la condition (2.8).

On note que le segment $[-n\bar{h}, \delta\bar{h}]$ de l'axe Ot est supposé dans la fermeture $\bar{\Omega}$ du domaine, ce qui est nécessairement vérifié pour h suffisamment petit. En effet, on fait l'hypothèse qu'un seul des segments joignant $\mathbf{x} = (x, y)$ et $(x \pm h, y)$ n'appartient pas à $\bar{\Omega}$. Si la frontière Γ coupe deux fois l'axe Ot de façon que les points d'intersection soient distants de moins de $(n+1)h$, alors la courbure de Γ est de l'ordre de $1/h$, ce qui contredit pour $h \to 0$ l'hypothèse $\Gamma \in C^2$.

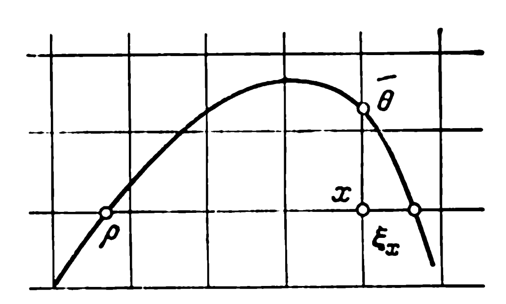


Fig. 4.2. Nœud irrégulier dans la direction x

Voyons le cas de la fig. 4.2. Les axes Ot et Ox ont même direction, et la frontière Γ coupe l'axe Ot la deuxième fois au point $\boldsymbol{\rho}$. On suppose qu'en coordonnées de l'axe Ot et de l'axe $O\sigma$ orthogonal parallèle à la direction de Oy et compté à partir du point $\mathbf{x}$, la courbe entre $\boldsymbol{\rho}$ et $\xi_\mathbf{x}$ est décrite par l'équation $\sigma = g(t)$. On note $\boldsymbol{\theta}$ le point d'intersection de l'axe $O\sigma$ et de Γ. On rappelle que $|\boldsymbol{\theta} - \mathbf{x}| \geqslant h$. Selon un théorème de Lagrange, il existe entre $\mathbf{x}$ et $\xi_\mathbf{x}$ de l'axe Ot un point t_1 tel que $g'(t_1) < -1$. Or, le segment $[\boldsymbol{\rho}, \xi_\mathbf{x}]$ renferme par

Rolle un point t_2 tel que $g'(t_2) = 0$. Le même résultat de Lagrange fait conclure à l'existence d'un point t_3 séparant t_1 et t_2 en lequel

$$g''(t_3) = \frac{g'(t_2) - g'(t_1)}{t_2 - t_1}.$$

Mais $|g'(t_2) - g'(t_1)| > 1$ et $|t_2 - t_1| \leqslant (n+1)h$, si bien que

$$|g''(t_3)| > \frac{1}{(n+1)h}.$$

On se place maintenant dans le cas où la frontière coupe deux segments joignant, par exemple, les nœuds $(x+h, y)$, $(x, y+h)$, d'une part, et $\mathbf{x} = (x, y)$, de l'autre. On fait prendre à l'axe Ot la

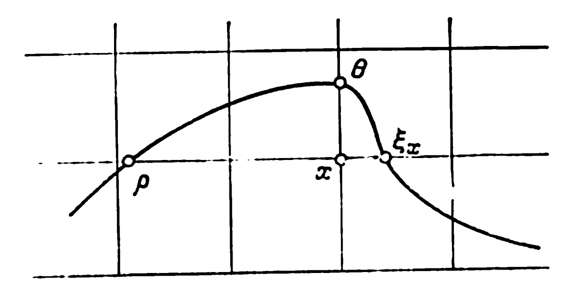


Fig. 4.3. Nœud irrégulier dans deux directions

direction du nœud frontière le plus proche de $\mathbf{x}$. Soit $\xi_{\mathbf{x}}$ ce point (fig. 4.3). La démonstration est répétée mot pour mot. On note que (2.8) confère au système d'équations algébriques la propriété d'être diagonalement dominant, la domination étant stricte aux points irréguliers. Avec ce résultat, on démontre la stabilité du problème aux différences (2.3), (2.10), ce qui conduit à son tour au

THÉORÈME 2.1 (voir [9]). *Soit, dans le problème* (2.1), (2.2),

$$\Gamma \in C^{2m+2+\lambda}, \qquad \varphi \in C^{2m+2+\lambda}(\Gamma), \qquad f \in C^{2m+\lambda}(\bar{\Omega})$$

pour m entier naturel et $\lambda \in (0, 1)$. *La solution* u^h *du système d'équations aux différences* (2.3), (2.10), $n = 2m$, *admet la représentation*

$$u^h = u + \sum_{k=1}^{m} h^{2k} w_k + h^{2m+\lambda} r_h \quad sur \quad \Omega_h. \tag{2.11}$$

Ici u *est la solution du problème* (2.1), (2.2), *les fonctions* w_k *sont dans* $C^{2m+2-2k+\lambda}(\bar{\Omega})$ *et indépendantes de* h, *et la fonction discrète* r_h *est bornée en valeur absolue par une constante indépendante de* h:

$$|r_h(\mathbf{x})| \leqslant c_2 \qquad \forall \mathbf{x} \in \Omega_h.$$

Les modalités d'emploi du développement (2.11) ne diffèrent en principe pas du cas unidimensionnel, mais l'interpolation devient plus ardue et on raffine de préférence avec les pas h, $h/2$, $h/3$, ... car cela évite d'interpoler au moins pour les nœuds de $\bar{\Omega}_h$.

L'approximation de la forme (2.10) présente deux inconvénients. Si l'on veut être dans la condition (2.8) pour $n > 2$, il faut prendre le pas $\bar{h}$ plusieurs fois plus grand que h. S'agissant de h faibles cela détériore sérieusement les propriétés d'approximation du polynôme de Lagrange (on voit augmenter la constante dans (2.6)), tandis que dans le cas de h importants, certains points d'interpolation sortent de Ω, si bien que l'approximation (2.10) n'est plus utile. On atteint plusieurs premiers ordres de précision moyennant d'autres approximations des conditions aux limites qui utilisent l'information sur l'équation (2.1) même. On garantit, par exemple, une solution extrapolée à l'ordre 3 en h avec l'approximation proposée dans [32], [128] et étudiée de ce point de vue dans [52]. La précision en h^4 est réalisée par les équations de [7], [13], [57], [70]. On trouve dans [11] une approximation qui conduit bien à une solution extrapolée en h^5, mais qui est basée sur des points très rapprochés. A la fin de ce paragraphe, on utilisera les nœuds irréguliers pour approcher l'équation de Poisson associée à un réseau de mailles non uniformes. Les équations aux différences correspondantes sont basées sur des points plus serrés que ceux de (2.10), et elles présentent l'avantage d'intervenir dans le schéma de décomposition pour l'équation de la chaleur. D'autre part, les approximations de la forme (2.10) s'appliquent telles quelles aux équations à coefficients variables, tandis que les autres schémas se compliquent singulièrement.

Le problème (2.3), (2.10) présente un autre défaut absent des schémas usuels à cinq points. C'est la non-symétrie de la matrice du système algébrique, qui rend inopérants plusieurs procédés de résolution efficaces. On évite cet inconvénient par passage au processus itératif

$$-w^k_{\bar{x}\hat{x}} - w^k_{\bar{y}\hat{y}} = f \quad \text{sur} \quad \Omega^r_h, \tag{2.12}$$
$$w^k(\mathbf{x}) = I^{2m}_\delta w^{k-1}(\mathbf{x}) + \lambda^{2m}_\delta \varphi(\xi_r), \qquad \mathbf{x} \in \Omega^{ir}_h.$$

disons, avec la condition initiale nulle $w^0 \equiv 0$. On résout à chaque pas un système de matrice définie positive symétrique. On montre que l'erreur diminue de moitié d'une itération à l'autre. On pose sur Ω_h

$$\psi^k = w^k - u^h,$$

avec ψ^k l'erreur sur la solution w^k, auquel cas

$$\begin{aligned} -\psi^k_{\bar{x}\hat{x}} - \psi^k_{\bar{y}\hat{y}} &= 0 \quad \text{sur} \quad \Omega^r_h, \\ \psi^k &= I^{2m}_\delta \psi^{k-1} \quad \text{sur} \quad \Omega^{ir}_h. \end{aligned} \tag{2.13}$$

On note que les égalités (2.13) vérifient le principe du maximum. Aussi $|\psi^k|$ atteint son maximum sur l'ensemble Ω_k^{ir}. Soit $\mathbf{x}_0 \in \Omega_k^{ir}$ ce point. On a

$$\max_{\Omega_k} |\psi^k| = |\psi^k(\mathbf{x}_0)| = |I_8^{2m}\psi^{k-1}(\mathbf{x}_0)|.$$

Par suite de la définition (2.7), on a la relation

$$\max_{\Omega_k} |\psi^k| \leqslant B(\delta) \max_{\Omega_k} |\psi^{k-1}|$$

qui se ramène à

$$\max_{\Omega_k} |\psi^k| \leqslant \frac{1}{2} \max_{\Omega_k} |\psi^{k-1}|$$

en vertu de (2.8). On note de plus qu'il ne faut pas chercher le plus exactement possible la solution de (2.12) pour chaque k. Appliquons à ce système un procédé itératif. On doit arrêter les itérations dès que la composante de plus grand module du résidu ξ^k relatif à la solution approchée $\tilde{w}^k$ du problème

$$-\tilde{w}^k_{\bar{x}x} - \tilde{w}^k_{\bar{y}y} = f + \xi^k \quad \text{sur} \quad \Omega_k^r,$$
$$\tilde{w}^k(\mathbf{x}) = I_8^{2m}\tilde{w}^{k-1}(\mathbf{x}) + \lambda_8^{2m}\varphi(\xi), \qquad \mathbf{x} \in \Omega_k^{ir}.$$

vérifie l'inégalité

$$\max_{\Omega_k^r} |\xi^k| \leqslant 2^{-k} \max_{\Omega_k^r} |\xi^0| = 2^{-k} \max_{\Omega_k^r} |f|.$$

Avec l'approximation $\tilde{w}^{k-1}$ suffisamment voisine, ce nombre d'itérations suffit pour trouver $\tilde{w}^k$.

4.2.2. Raffinement par différences d'ordre supérieur

Nous allons décrire un procédé de raffinement de la solution approchée sur le réseau Ω_k même. L'idée en appartient à Fox [86], et c'est Volkov qui a été le premier à le justifier pour des équations aux dérivées partielles [6].

Conformément à [9], on désigne par $D_x^{(p,n)}$ l'opérateur aux différences du calcul approché de la dérivée p-ième par rapport à x d'une fonction définie sur le réseau avec h constant de l'axe Ox. L'opérateur $D_x^{(p,n)}$ est obtenu en dérivant p fois la formule d'interpolation de Newton par différences d'ordre n inclus (voir [67]). On note que les opérateurs

$$D_x^{(m)} = \sum_{k=2}^{m+1} \frac{2h^{2k-2}}{(2k)!} D_x^{(2k,\,2m+2)}, \qquad D_y^{(m)} = \sum_{k=2}^{m+1} \frac{2h^{2k-2}}{(2k)!} D_y^{(2k,\,2m+2)} \qquad (2.14)$$

donnent les dérivées secondes avec une précision en $h^{2m+\lambda}$ pour toute fonction $\psi \in C^{2m+2+\lambda}(\bar\Omega)$, par exemple

$$-\Delta\psi = -\psi_{\bar{x}x} - \psi_{\bar{y}y} + D_x^{(m)}\psi + D_y^{(m)}\psi + h^{2m+\lambda}\theta_3. \qquad (2.15)$$

où

$$|\theta_3(\mathbf{x})| \leqslant c_3 \|\psi\|_{C^{2m+2+\lambda}(\bar\Omega)}$$

à condition que tous les points utilisés par $D_x^{(m)}$ et $D_y^{(m)}$ appartiennent à $\bar\Omega$. Pour qu'il en soit ainsi, on fait agir diverses modifications de $D_x^{(m)}$ et $D_y^{(m)}$ au voisinage de la frontière Γ et à une certaine distance d'elle. On se sert si possible des différences centrales seules. Si l'on introduit les opérateurs aux différences

$$\Delta f(x) = f(x) - f(x-h),$$
$$\nabla f(x) = f(x+h) - f(x),$$
$$\square f(x) = f(x+h/2) - f(x-h/2)$$

et si l'on établit de façon récurrente que

$$\nabla^k f = \nabla(\nabla^{k-1} f)$$

(on obtient la même chose pour Δ^k et $\square^k$), on a alors la liste suivante des opérateurs $D_x^{(m)}$ disponibles :

$$D_x^{(m)} = \frac{\Delta^4}{12} - \frac{\Delta^5}{12} + \frac{13}{180}\Delta^6 - \frac{11}{180}\Delta^7 + \frac{87}{1680}\Delta^8 + \ldots,$$

$$D_x^{(m)} = \nabla\left(\frac{\Delta^3}{12} - \frac{\Delta^5}{90} + \frac{\Delta^6}{90} - \frac{47}{5040}\Delta^7\right) + \ldots,$$

$$D_x^{(m)} = \nabla^2\left(\frac{\Delta^2}{12} + \frac{\Delta^3}{12} - \frac{\Delta^4}{90} + \frac{1}{560}\Delta^6\right) + \ldots,$$

. . . .

$$D_x^{(m)} = \frac{\square^4}{12} - \frac{\square^6}{90} + \frac{\square^8}{560} - \ldots,$$

$$D_x^{(m)} = \Delta^2\left(\frac{\nabla^2}{12} - \frac{\nabla^3}{12} - \frac{\nabla^4}{90} + \frac{1}{560}\Delta^6\right) + \ldots.$$

$$D_x^{(m)} = \Delta\left(\frac{\nabla^3}{12} - \frac{\nabla^5}{90} - \frac{\nabla^6}{90} - \frac{47}{5040}\nabla^7\right) + \ldots,$$

$$D_x^{(m)} = \frac{\nabla^4}{12} + \frac{\nabla^5}{12} + \frac{13}{180}\nabla^6 + \frac{11}{180}\nabla^7 + \frac{87}{1680}\nabla^8 + \ldots$$

Il y a lieu de dire qu'avec h suffisamment petit et m fini, on écrit pour un nœud régulier une modification de $D_x^{(m)}$ et $D_y^{(m)}$ telle que tous les nœuds utilisés soient dans $\bar{\Omega}$.

Soit, par exemple, le cas de la fig. 4.4. Le point (x, y) est régulier, si bien que la distance de (x, y) et du point frontière (x, η) est supérieure à h. Si la distance séparant les nœuds frontières (θ_1, y) et (θ_2, y) est plus petite que nh, il existe par Lagrange sur l'arc compris entre (θ_1, y) et (x, η) un point tel que la dérivée en ce point soit supérieure à $1/n$ et sur l'arc d'extrémités (θ_2, y), (x, η) un point en lequel la dérivée est inférieure à $-1/n$. On trouve selon le même théorème un point où la courbure est en $O(h^{-1})$. Or, cela contredit l'hypothèse de h suffisamment petits.

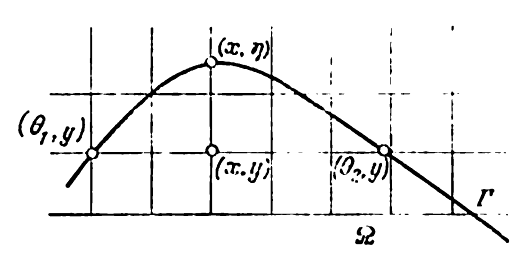


Fig. 4.4. Position possible du nœud régulier (x, y)

On énonce le problème aux différences

$$-u^h_{\bar{x}x} - u^h_{\bar{y}y} = f - D_x^{(m)} u^h - D_y^{(m)} u^h \quad \text{sur} \quad \Omega'_h$$

$$u^h(\mathbf{x}) = l_\delta^{(2m)} u^h(\mathbf{x}) + \lambda_\delta^{2m} \varphi(\xi_x), \qquad \mathbf{x} \in \Omega''_h.$$

Ses erreurs d'approximation sont en $h^{2m+\lambda}$ pour $u \in C^{2m+2+\lambda}(\bar{\Omega})$. Le problème de la stabilité reste cependant ouvert, et le schéma aboutit des fois à une précision mauvaise.

On remplace donc ce schéma par une succession de problèmes aux différences

$$-v^q_{\bar{x}x} - v^q_{\bar{y}y} = f - D_x^{(m)} v^{q-1} - D_y^{(m)} v^{q-1} \quad \text{sur} \quad \Omega'_h, \tag{2.16}$$

$$v^q(\mathbf{x}) = l_\delta^{2m} v^q(\mathbf{x}) + \lambda_\delta^{2m} \varphi(\xi_x), \qquad \mathbf{x} \in \Omega''_h, \tag{2.17}$$

$$q = 1, 2, \ldots, m+1.$$

On initialise avec l'approximation $v^0 \equiv 0$. Alors v^1 coïncide évidemment avec la solution du problème (2.3), (2.10), et elle ne diffère de la solution du problème différentiel que par une quantité en h^2. Les solutions suivantes v^q sont de plus en plus exactes, l'ordre de précision étant limité par la régularité seule des données du problème différentiel.

THÉORÈME 2.2 (voir [10]). *On suppose que le problème* (2.1), (2.2) *remplit les conditions*

$$\Gamma \in C^{2m+2+\lambda}, \qquad \varphi \in C^{2m+2+\lambda}(\Gamma), \qquad f \in C^{2m+\lambda}(\bar{\Omega}),$$

avec $m \geq 1$ *et* $\lambda \in (0, 1)$. *La solution* v^{m+1} *du problème* (2.16), (2.17) *admet la représentation*

$$v^{m+1} = u + h^{2m+\lambda} \left| \ln \frac{2b}{h} \right|^m r_h \quad sur \quad \Omega_h. \qquad (2.18)$$

Ici u *est la solution du problème* (2.1), (2.2) *et* r_h *une fonction discrète bornée :*

$$| r_h (\mathbf{x}) | \leq c_3 \qquad \forall \mathbf{x} \in \Omega_h. \qquad (2.19)$$

La méthode (2.16), (2.17) se simplifie notablement si l'on remplace (2.17) par la relation

$$v^q (\mathbf{x}) = l_\delta^{2m} v^{q-1} (\mathbf{x}) + \lambda_\delta^{2m} \varphi (\xi_{\mathbf{x}}), \qquad \mathbf{x} \in \Omega_h^{ir}.$$

En effet, le problème devient auto-adjoint pour toute fonction v^q. D'autre part, v^q n'approche plus u à l'ordre h^{2q}. Cet artifice est néanmoins à recommander lorsqu'on fait des itérations pour q fixe. On en a d'ailleurs parlé au numéro précédent.

4.2.3. Approximation à plusieurs points des dérivées secondes

On va examiner plusieurs façons d'approcher les dérivées secondes, et on se propose d'élaborer des procédés qui garantissent une forme spéciale de l'erreur d'approximation pour les nœuds irréguliers.

Soit $\psi (t)$ une fonction de $C^{m+\alpha} [-1, 1]$, avec $m \geq 2$, $\alpha \in (0, 1)$. On rappelle que si l'on utilise trois valeurs et les points équidistants $-h$, 0, h, l'erreur d'approximation sur la dérivée seconde est

$$\psi_{\bar{t}t} (0) - \psi'' (0) = \sum_{k=1}^{[(m-2)/2]} h^{2k} \frac{2 \psi^{(2k+2)}(0)}{(2k+2)!} + h^{m+\alpha} \theta, \qquad (2.20)$$

où

$$| \theta | \leq \frac{2}{m!} \| \psi \|_{C^{m+\alpha}[-1,1]}. \qquad (2.21)$$

Nous voulons obtenir une erreur d'approximation de la même forme pour des points non régulièrement espacés. On se donne, sur l'axe Ot, $n + 2$ points $\delta h, 0, -h, \ldots, -nh$, où $\delta \in (0, 4/3)$. On calcule $\psi(h)$ moyennant le polynôme d'interpolation de Lagrange :

$$\psi (h) = \sum_{k=0}^{n} a_{n,k} \psi (- kh) + a_{n,\delta} \psi (\delta h) + r^h, \qquad (2.22)$$

où

$$a_{n,k} = \frac{(-1)^{k+1} (n+1)! (1-\delta)}{(k+1)! (k+\delta) (n-k)!}, \qquad a_{n,\delta} = \prod_{k=0}^{n} \frac{k+1}{k+\delta},$$

Si $m \geqslant n+2$, alors le reste est évalué comme suit (voir [67]):

$$|r^h| \leqslant \frac{|h-\delta h|\, h(h+h)\ldots(h+nh)}{(n+2)!}\max_{[-1,1]}|\psi^{(n+2)}| \leqslant$$

$$\leqslant h^{n+2}\frac{|1-\delta|}{n+2}\max_{[-1,1]}|\psi^{(n+2)}|. \qquad (2.23)$$

Si $m < n+2$, i.e. la fonction ψ est insuffisamment régulière, le reste est une quantité d'ordre supérieur. On démontre ce résultat. Soit le polynôme

$$\varphi(t) = \sum_{i=0}^{m} \alpha_i t^i.$$

Comme sa dérivée $n+2$-ième est identiquement nulle, la relation (2.22) entraîne pour $\varphi(h)$

$$\varphi(h) = \sum_{k=0}^{n} a_{n,k}\,\varphi(-kh) + a_{n,\delta}\,\varphi(\delta h). \qquad (2.24)$$

Revenons à la fonction ψ. Le développement de Taylor implique

$$\psi(-kh) = \sum_{i=0}^{m} h^i \frac{(-k)^i}{i!}\psi^{(i)}(0) + \frac{h^{m+\alpha}|k|^{m+\alpha}}{m!}\theta_k, \qquad (2.25)$$

$$k = -1,\ 0,\ \ldots,\ n,$$

et

$$\psi(\delta h) = \sum_{i=0}^{m} h^i \frac{\delta^i}{i!}\psi^{(i)}(0) + \frac{h^{m+\alpha}\delta^{m+\alpha}}{m!}\theta_\delta.$$

avec

$$\max_{-1\leqslant k\leqslant n}\{|\theta_k|, |\theta_\delta|\} \leqslant \|\psi\|_{C^{m+\alpha}[-1,1]}.$$

On pose les coefficients α_i du polynôme φ égaux à $\psi^{(i)}(0)/i!$, auquel cas les développements (2.25) s'écrivent

$$\psi(-kh) = \varphi(-kh) + \frac{h^{m+\alpha}|k|^{m+\alpha}}{m!}\theta_k.$$

$$\psi(\delta h) = \varphi(\delta h) + \frac{h^{m+\alpha}\delta^{m+\alpha}}{m!}\theta_\sigma.$$

On les porte dans (2.22), il vient par suite de (2.24)

$$\frac{h^{m+\alpha}}{m!}\theta_1 = \frac{h^{m+\alpha}}{m!}\sum_{k=1}^{n} a_{n,k}\,k^{m+\alpha}\theta_k + a_{n,\delta}\frac{h^{m+\alpha}\delta^{m+\alpha}}{m!}\theta_\delta + r^h.$$

D'où

$$|r^h| \leqslant \frac{h^{m+\alpha}}{m!} \|\psi\|_{C^{m+\alpha}[-1,1]} \left(1 + \sum_{k=1}^{n} |a_{n,k}| k^{m+\alpha} + a_{n,\delta} \delta^{m+\alpha}\right).$$

Puisque

$$|a_{n,k}| \leqslant (n+1)! \quad \text{pour} \quad k = 1, \ldots, n, \qquad \delta \in (0,3)$$

et

$$\delta a_{n,\delta} \leqslant n + 1$$

on a

$$r^h| \leqslant | \frac{h^{m+\alpha}}{m!} \|\psi\|_{C^{m+\alpha}[-1,1]} (1 + n(n+1)!\ n^{m+\alpha} + (n+1)2^{m+\alpha-1}).$$

On pose

$$c_4 - \frac{1}{m!} \|\psi\|_{C^{m+\alpha}[-1,1]} (1 + (n+1)[(n+1)!n^{m+\alpha} + 2^m]).$$

d'où l'estimation

$$|r^h| \leqslant c_4 h^{m+\alpha}, \qquad m < n + 2. \tag{2.26}$$

On chasse le reste r^h de (2.22), il vient la valeur approchée

$$\tilde{\psi}(h) - \sum_{k=0}^{n} a_{n,k}\ \psi(-kh) + a_{n,\delta}\ \psi(\delta h). \tag{2.27}$$

On approche la dérivée seconde :

$$\tilde{\psi}_{ll}(0) \equiv \frac{\tilde{\psi}(h) - 2\psi(0) + \psi(-h)}{h^2} =$$

$$= \sum_{k=0}^{n} b_{n,k}\ \psi(-kh) + b_{n,\delta}\psi(\delta h). \tag{2.28}$$

où

$$b_{n,0} = \frac{a_{n,0} - 2}{h^2}, \qquad b_{n,1} - \frac{a_{n,1} + 1}{h^2}, \qquad b_{n,k} = \frac{a_{n,k}}{h^2}, \qquad b_{n,\delta} - \frac{a_{n,\delta}}{h^2},$$

$$k = 2, \ldots, n.$$

Cette dérivée satisfait aux égalités

$$\tilde{\psi}_{ll}(0) - \psi_{ll}(0) - \frac{\tilde{\psi}(h) - \psi(h)}{h^2} = \frac{r^h}{h^2}.$$

Aussi le développement (2.20) et les estimations (2.21), (2.23), (2.26) conduisent à

$$\tilde{\psi}_{\bar{x}x}(0) - \psi''(0) = \sum_{k=1}^{s} \frac{2h^{2k}}{(2k+2)!}\,\psi^{(2k+2)}(0) + r_1^h, \qquad (2.29)$$

avec

$$s = \left[\frac{1}{2}\min\{n-1,\ m-2\}\right],$$

le reste étant évalué comme suit:

$$|r_1^h| \leq \begin{cases} c_5 h^n & \text{si} \quad n+2 \leq m, \\ c_6 h^{m-2+\alpha} & \text{si} \quad 1 \leq m \leq n+1, \end{cases} \qquad (2.30)$$

où les constantes c_5, c_6 ne dépendent pas de h et δ.

Lorsque $n = 1, 2, 3$, les coefficients $b_{n,k}$ jouissent de la propriété importante

$$\frac{-b_{n,0} - \sum\limits_{k=1}^{n} |b_{n,k}|}{\dfrac{2}{h^2} + \sum\limits_{k=1}^{n} |b_{n,k}|} \geqslant c_7 > 0, \qquad (2.31)$$

où la constante c_7 indépendante de h et δ varie avec n. On démontre cette inégalité.

Soit $n = 1$, auquel cas

$$b_{1,\delta} = \frac{2}{\delta(\delta+1)h^2}, \qquad b_{1,0} = -\frac{2}{\delta h^2}, \qquad b_{1,1} = \frac{2}{(1+\delta)h^2}.$$

Le premier membre de (2.31) vaut

$$\frac{\dfrac{2}{\delta h^2} - \dfrac{2}{(1+\delta)h^2}}{\dfrac{2}{h^2} + \dfrac{2}{(1+\delta)h^2}} = \frac{1}{2\delta + \delta^2}.$$

Comme $\delta \in (0, 4/3)$, on a

$$\frac{1}{2\delta + \delta^2} \geqslant \frac{9}{40},$$

et on pose $c_7 = 9/40$ pour $n = 1$.

Soit $n = 2$. Alors

$$b_{2,\delta} = \frac{6}{\delta(\delta+1)(\delta+2)h^2}, \qquad b_{2,0} = -\frac{3-\delta}{\delta h^2},$$

$$b_{2,1} = \frac{4-2\delta}{(1+\delta)h^2}, \qquad b_{2,2} = -\frac{1-\delta}{(2+\delta)h^2}.$$

On se place dans le cas $\delta \in (0, 1]$. Le dénominateur dans le premier membre de (2.31) étant positif, on a

$$\frac{\dfrac{3-\delta}{\delta h^2} - \dfrac{4-2\delta}{(1+\delta)h^2} - \dfrac{1-\delta}{(2+\delta)h^2}}{\dfrac{2}{h^2} + \dfrac{4-2\delta}{(1+\delta)h^2} + \dfrac{1-\delta}{(2+\delta)h^2}} \geqslant \frac{\dfrac{3-\delta}{\delta} - \dfrac{4-2\delta}{1+\delta} - \dfrac{1-\delta}{2}}{2 + \dfrac{4-2\delta}{1+\delta} + \dfrac{1-\delta}{2}} =$$

$$= \frac{6 - 5\delta + 2\delta^2 + \delta^3}{\delta(13-\delta^2)} \geqslant \frac{6 - 5\delta + 2\delta^2}{13}.$$

L'expression $6 - 5\delta + 2\delta^2$ atteint son minimum sur $(0, 1]$ pour $\delta = 1$, si bien que

$$\frac{6 - 5\delta + 2\delta^2}{13} \geqslant \frac{3}{13}.$$

Soit maintenant le cas $\delta \in (1, 4/3]$. On se rappelle la positivité du dénominateur du premier membre de (2.31) et on obtient

$$\frac{\dfrac{3-\delta}{\delta h^3} - \dfrac{4-2\delta}{(1+\delta)h^2} + \dfrac{1-\delta}{(2+\delta)h^3}}{\dfrac{2}{h^2} + \dfrac{4-2\delta}{(1+\delta)h^2} - \dfrac{1-\delta}{(2+\delta)h^2}} \geqslant \frac{\dfrac{3-\delta}{\delta} - \dfrac{4-2\delta}{1+\delta} + \dfrac{1-\delta}{3}}{2 + \dfrac{4-2\delta}{1+\delta} - \dfrac{1-\delta}{3}} =$$

$$= \frac{9 - 5\delta + 3\delta^2 - \delta^3}{17\delta + \delta^3}.$$

Dans la dernière fraction, le numérateur est monotone décroissant et le dénominateur est monotone croissant avec δ tendant vers 0 par valeurs positives. On pose $\delta = 4/3$ et on a donc

$$\frac{9 - 5\delta + 3\delta^2 - \delta^3}{17\delta + \delta^3} \geqslant \frac{143}{676} \geqslant \frac{1}{5}.$$

Ainsi, on réunit les deux cas et on pose $c_7 = 1/5$ pour $n = 2$.

Quelle est la constante c_7 pour $n = 3$? On écrit les coefficients

$$b_{3,\delta} = \frac{24}{\delta(1+\delta)(2+\delta)(3+\delta)h^2}, \qquad b_{3,0} = -\frac{2(2-\delta)}{\delta h^2},$$

$$b_{3,1} = \frac{7-5\delta}{(1+\delta)h^2}, \qquad b_{3,2} = \frac{-4(1-\delta)}{(2+\delta)h^2}, \quad b_{3,3} = \frac{1-\delta}{(3+\delta)h^2}.$$

Soit le cas $\delta \in (0, 1]$. On calcule le premier membre de (2.31):

$$\frac{\dfrac{2(2-\delta)}{\delta h^2} - \dfrac{7-5\delta}{(1+\delta)h^2} - \dfrac{4(1-\delta)}{(2+\delta)h^2} - \dfrac{1-\delta}{(3+\delta)h^2}}{\dfrac{2}{h^2} + \dfrac{7-5\delta}{(1+\delta)h^2} + \dfrac{4(1-\delta)}{(2+\delta)h^2} + \dfrac{1-\delta}{(3+\delta)h^2}} \geqslant$$

$$\geqslant \frac{\dfrac{2(2-\delta)}{\delta} - \dfrac{7-5\delta}{1+\delta} - \dfrac{4(1-\delta)}{2+\delta} - \dfrac{1}{3}}{2 + \dfrac{7-5\delta}{1+\delta} + \dfrac{4(1-\delta)}{2+\delta} + \dfrac{1}{3}} = \frac{24 - 32\delta + 20\delta^3}{68\delta + 12\delta^2 - 20\delta^3}.$$

On trouve par certaines transformations

$$\frac{24 - 32\delta + 20\delta^3}{68\delta + 12\delta^2 - 20\delta^3} \geqslant \frac{6 - 8\delta + 8\delta^3/3}{17\delta + 3\delta^2}.$$

Le numérateur de la dernière fraction décroît monotonément sur $(0, 1]$, tandis que le dénominateur est monotone croissant. On pose $\delta = 1$, il vient

$$\frac{6 - 8\delta + 8\delta^3/3}{17\delta + 3\delta^2} = \frac{1}{30}.$$

Soit enfin le cas $\delta \in (1, 4/3]$. Le premier membre de (2.31) s'écrit

$$\frac{\dfrac{2(2-\delta)}{\delta h^2} - \dfrac{7-5\delta}{(1+\delta)h^2} + \dfrac{4(1-\delta)}{(2+\delta)h^2} + \dfrac{1-\delta}{(3+\delta)h^2}}{\dfrac{2}{h^2} + \dfrac{7-5\delta}{(1+\delta)h^2} - \dfrac{4(1-\delta)}{(2+\delta)h^2} - \dfrac{1-\delta}{(3+\delta)h^2}} \geqslant$$

$$\geqslant \frac{\dfrac{2(2-\delta)}{\delta} - \dfrac{7-5\delta}{1+\delta} + \dfrac{4(1-4/3)}{3} + \dfrac{1-4/3}{4}}{2 + \dfrac{7-5\delta}{1+\delta} - \dfrac{4(1-4/3)}{3} - \dfrac{1-4/3}{4}} = \frac{\dfrac{4-5\delta+3\delta^2}{\delta(1+\delta)} - \dfrac{19}{36}}{2 + \dfrac{7-5\delta}{1+\delta} + \dfrac{19}{36}}.$$

Le polynôme $4 - 5\delta + 3\delta^2$ croît sur $(1, 4/3]$, si bien qu'on évalue le numérateur par

$$\frac{4 - 5\delta + 3\delta^2}{\delta(1+\delta)} - \frac{19}{36} \geqslant \frac{2}{\dfrac{4}{3}\left(1 + \dfrac{4}{3}\right)} - \frac{19}{36} \geqslant \frac{1}{2}.$$

Quant au dénominateur, on a

$$2 + \frac{7-5\delta}{1+\delta} + \frac{19}{36} \leqslant 2 + \frac{2}{2} + \frac{19}{36} = \frac{127}{36}.$$

Ainsi, toute la fraction est bornée inférieurement par 18/127. On réunit les deux cas et on a $c_7 = 1/30$ pour $n = 3$. On a calculé c_7

d'une manière assez grossière car notre seul but a été d'en démontrer l'existence.

La propriété (2.31) garantit, on le verra plus loin, la stabilité du schéma aux différences. On ne saurait choisir c_7 positive pour $n \geqslant 4$ parce que le premier membre de (2.31) devient négatif pour certains $\delta \in (0, 4/3]$.

On se propose d'obtenir une estimation de la forme (2.31) (qui garantira plus loin la stabilité du schéma aux différences). On effectue l'interpolation de pas $\bar{h}$ multiple de h (voir le début du paragraphe) : on pose $\bar{h} = ph$, avec p entier naturel et on répète les raisonnements (2.22) à (2.27), il vient la formule

$$\tilde{\psi}_h(h) = \sum_{k=0}^{n} \alpha_{n,k}\, \psi(-kph) + \alpha_{n,\delta}\, \psi(\delta h), \tag{2.32}$$

où

$$\begin{aligned} \alpha_{n,k} &= (-1)^{k+1} \frac{(1-\delta)(p+1)(2p+1)\dots(np+1)}{(kp+1)(kp+\delta)k!(n-k)!p^n}, \\ \alpha_{n,\delta} &= \frac{(p+1)(2p+1)\dots(np+1)}{\delta(p+\delta)(2p+\delta)\dots(np+\delta)}. \end{aligned} \tag{2.33}$$

L'écart entre cette valeur approchée et la valeur exacte est la quantité

$$r_2^h = \tilde{\psi}(h) - \psi(h)$$

qui dépend de la relation entre m et n. Si $m \geqslant n + 2$, alors

$$|r_2^h| \leqslant h^{n+2} \frac{|1-\delta|(p+1)(2p+1)\dots(np+1)}{(n+2)!} \max_{[-1,1]} |\psi^{(n+2)}| \leqslant$$
$$\leqslant h^{n+2} p^n \frac{|1-\delta|}{n+2} \max_{[-1,1]} |\psi^{(n+2)}|.$$

Si $m \leqslant n + 1$, on a

$$|r_2^h| \leqslant \left(\frac{h^{m+\alpha}}{m!} + \sum_{k=1}^{n} |\alpha_{n,k}| \frac{h^{m+\alpha}(kp)^{m+\alpha}}{m!} + |\alpha_{n,\delta}| \frac{h^{m+\alpha}\delta^{m+\alpha}}{m!} \right) \|\psi\|_{C^{m+\alpha}[-1,1]} \tag{2.34}$$

Les égalités (2.33) entraînent

$$|\alpha_{n,k}| \leqslant \frac{|1-\delta|(n+1)!}{k^2 p^2 k!(n-k)!} \leqslant \frac{|1-\delta|(n+1)!}{p^2}, \qquad k = 1, \dots, n, \tag{2.35}$$
$$\delta |\alpha_{n,k}| \leqslant n + 1.$$

Donc

$$|r_2^h| \leqslant \frac{h^{m+\alpha}}{m} \| \psi \|_{C^{m+\alpha}[-1,1]} \times$$

$$\times \left(1 + n\,(np)^{m+\alpha} \frac{|1-\delta|\,(n+1)!}{p^2} + (n+1)\,\delta^{m+\alpha-1}\right).$$

Etant donné que $\delta \in (0, 4/3]$, on a la majoration

$$|r_2^h| \leqslant c_8\, h^{m+\alpha}, \qquad m \leqslant n+1, \tag{2.36}$$

avec la constante c_8 indépendante de p.

On utilise (2.32) pour approcher la dérivée seconde:

$$\tilde{\psi}_{\bar{t}t}(0) = \frac{\tilde{\psi}(h) - 2\psi(0) + \psi(-h)}{h^2} =$$

$$= \sum_{i=0}^{np} \beta_{n,i}\, \psi(-ih) + \beta_{n,\delta}\, \psi(\delta h), \tag{2.37}$$

où

$$\beta_{n,\delta} = \frac{\alpha_{n,\delta}}{h^2}, \qquad \beta_{n,0} = \frac{\alpha_{n,0} - 2}{h^2},$$

$$\beta_{n,1} = \begin{cases} 1/h^2 & \text{si } p > 1, \\ (\alpha_{n,1} + 1)/h^2 & \text{si } p = 1, \end{cases}$$

$$\beta_{n,i} = \begin{cases} \alpha_{n,k}/h^2 & \text{si } i = pk, \\ 0 & \text{si } i = 2, \ldots, np;\ i \neq pk. \end{cases} \tag{2.38}$$

La relation (2.37) généralise la formule (2.28) pour p entier naturel. On choisit p de façon qu'on ait

$$\frac{-\beta_{n,0} - \sum\limits_{k=1}^{np} |\beta_{n,k}|}{\dfrac{2}{h^2} + \sum\limits_{k=1}^{np} |\beta_{n,k}|} \geqslant c_7, \tag{2.39}$$

où la constante positive c_7 est indépendante de $\delta \in (0, 4/3]$. On y arrive nécessairement pour p suffisamment grand et h assez petit. En effet, les estimations (2.35) et les égalités (2.38) impliquent pour $p \to \infty$

$$\beta_{n,0} \to -\frac{2}{h^2}, \qquad \beta_{n,1} \to \frac{1}{h^2}, \qquad \beta_{n,i} \to 0,$$
$$i = 2, \ldots, np.$$

On note que quand $n = 4$ ou $n = 5$ (qui garantissent finalement la précision en h^{n+2} de la solution approchée), on choisit p déterminés par (2.39) relativement faibles ($p = 2$ ou $p = 3$ respectivement).

Une fois p fixé, on utilise pour $\tilde{\psi}_{tt}(0) - \psi_{tt}(0)$ les estimations (2.34) et (2.36). Il est clair que l'approximation (2.37) admet le développement (2.29) dont le reste est évalué par (2.30).

On construit le problème aux différences. On suppose que le domaine Ω est inclus dans le carré $\{-b < x < b, -b < y < b\}$. On couvre le carré d'un réseau carré de pas $h = b/N$ formé par les droites $x_i = ih$ et $y_j = jh$, où $i, j = -N, \ldots, N$. Un nœud $\mathbf{x} \in \Omega_h$ est dit *régulier dans la direction* x si $\bar{\Omega}$ contient le segment d'extrémités $(x - h, y)$ et $(x + h, y)$. On définit de même un nœud régulier dans la direction y. Tous les nœuds réguliers dans la direction x et tous les nœuds réguliers dans la direction y forment les ensembles respectifs $\Omega^r_{h,x}$ et $\Omega^r_{h,y}$. On dit naturellement que

$$\Omega^{ir}_{h,x} = \Omega_h \setminus \Omega^r_{h,x}$$

constitue l'ensemble des nœuds *irréguliers dans la direction* x et que

$$\Omega^{ir}_{h,y} = \Omega_h \setminus \Omega^r_{h,y}$$

est l'ensemble des nœuds irréguliers dans la direction y. Il existe évidemment entre ces ensembles et Ω^r_h, Ω^{ir}_h introduits plus haut les relations

$$\Omega^r_h = \Omega^r_{h,x} \cap \Omega^r_{h,y}, \qquad \Omega^{ir}_h = \Omega^{ir}_{h,x} \cup \Omega^{ir}_{h,y}.$$

La dérivée $\partial^2 u/\partial x^2$ est approchée en chaque nœud $(x, y) \in \Omega^r_{h,x}$ par la différence seconde $u^h_{\bar{x}x}(x, y)$, et elle l'est en chaque $(x, y) \in \Omega^{ir}_{h,x}$ par les relations établies plus haut. On procède de même en ce qui concerne $\partial^2 u/\partial y^2$.

On choisit le nombre n caractéristique de l'ordre d'approximation aux nœuds irréguliers. On suppose que le problème différentiel vérifie les hypothèses du théorème 1.2 pour un entier $l \geqslant 2$ et $\alpha \in (0, 1)$. On pose $n = l - 1$.

Soit $\mathbf{x} \in \Omega^{ir}_{h,x}$, auquel cas on trouve dans la direction parallèle à l'axe Ox un point $\xi_x \in \Gamma_{h,x}$ qui soit le plus proche de $\mathbf{x}$. On confond l'origine de l'axe Ot avec le point $\mathbf{x}$, on dirige l'axe de $\mathbf{x}$ vers ξ_x et on utilise l'approximation (2.37) de la dérivée seconde $\partial^2 u/\partial x^2(\mathbf{x})$:

$$\tilde{u}_{tt}(\mathbf{x}) = J^x_{l-1}\, u(\mathbf{x}) + \rho^x_{l-1}\, \varphi(\xi_x), \tag{2.40}$$

où

$$J^x_n\, u(\mathbf{x}) = \sum_{i=0}^{np} \beta^x_{n,i}\, u(x \pm ih, y), \qquad \rho^x_n = \beta^x_{n,\delta}\,,$$

$\delta^x = |\xi_x - \mathbf{x}|$ et on prend le signe plus ou le signe moins selon que les axes Ot et Ox ont, oui ou non, même direction.

REMARQUE. Il se peut que tous les points utilisés par l'opérateur J_{l-1}^{x} ne soient pas dans $\bar{\Omega}$ (fig. 4.5), si bien que l'approximation (2.40) n'a pas de sens. Dans ce cas, on élimine de Ω_h les points symbolisés par ✱. Les points ○ deviennent irréguliers dans la direction y. La distance de ces points et de la frontière Γ est supérieure à h, et l'écart est en h^2 par suite de la régularité de Γ. On comprend désormais pourquoi nous avons testé les propriétés (2.31), (2.39) non seulement pour $\delta \in (0, 1)$, mais encore pour δ plus grands que un, à savoir $\delta \in (1, 4/3]$.

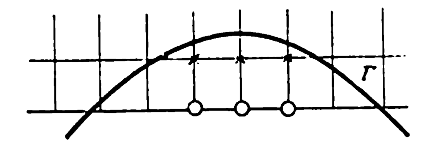


Fig. 4.5. Changements effectués dans les domaines de discrétisation dans le cas où la position du nœud x est mauvaise
✱ — points éliminés de Ω_h; ○ — points devenus irréguliers dans la direction y.

On traite de même le cas $\mathbf{x} \in \Omega_{h,\,y}^{r}$. La dérivée $\partial^2 u/\partial y^2$ est approchée par la dérivée aux différences $u_{\bar{y}\bar{y}}(\mathbf{x})$. Si $\mathbf{x} \in \Omega_{h,\,y}^{ir}$, alors on trouve dans la direction parallèle à l'axe Oy un point $\boldsymbol{\eta}_{\mathbf{x}} \in \Gamma_{h,\,y}$ qui soit le plus proche de $\mathbf{x}$. On compte l'axe Ot à partir de $\mathbf{x}$ et on le dirige de $\mathbf{x}$ vers $\boldsymbol{\eta}_{\mathbf{x}}$. L'approximation (2.37) de la dérivée seconde $\partial^2 u/\partial y^2$ s'écrit

$$\tilde{u}_{tt}(\mathbf{x}) = J_{l-1}^{y}\, u(\mathbf{x}) + \rho_{l-1}^{y}\, \varphi(\boldsymbol{\eta}_{\mathbf{x}}), \tag{2.41}$$

où

$$J_n^{y}\, u(\mathbf{x}) = \sum_{i=0}^{np} \beta_{n,\,i}^{y}\, u(x, y \pm ih), \qquad \rho_{l-1}^{y} = \beta_{n,\,\delta}^{y},$$

$\delta^{y} = |\boldsymbol{\eta}_{\mathbf{x}} - \mathbf{x}|$, et on prend le signe plus ou le signe moins selon que la direction de l'axe Ot est, oui ou non, celle de l'axe Oy. S'agissant de l'opérateur J_{l-1}^{y} et de l'approximation (2.41), la remarque ci-dessus reste en vigueur.

Ainsi, il correspond aux nœuds du réseau Ω_h quatre types d'approximations de l'équation (2.1):

$$-u_{\bar{x}\bar{x}}^{h}(\mathbf{x}) - u_{\bar{y}\bar{y}}^{h}(\mathbf{x}) = f(\mathbf{x}), \tag{2.42}$$
$$\mathbf{x} \in \Omega_h^{r};$$

$$-J_{l-1}^{x}\, u^{h}(\mathbf{x}) - u_{\bar{y}\bar{y}}^{h}(\mathbf{x}) = f(\mathbf{x}) + \rho_{l-1}^{x}\, \varphi(\boldsymbol{\xi}_{\mathbf{x}}), \tag{2.43}$$
$$\mathbf{x} \in \Omega_{h,\,x}^{ir} \cap \Omega_{h,\,y}^{r};$$

$$-u_{\bar{x}\bar{x}}^{h}(\mathbf{x}) - J_{l-1}^{y}\, u^{h}(\mathbf{x}) = f(\mathbf{x}) + \rho_{l-1}^{y}\, \varphi(\boldsymbol{\eta}_{\mathbf{x}}), \tag{2.44}$$
$$\mathbf{x} \in \Omega_{h,\,x}^{r} \cap \Omega_{h,\,y}^{ir};$$

$$-J_{l-1}^{x}\, u^{h}(\mathbf{x}) - J_{l-1}^{y}\, u^{h}(\mathbf{x}) = f(\mathbf{x}) + \rho_{l-1}^{x}\, \varphi(\boldsymbol{\xi}_{\mathbf{x}}) + \rho_{l-1}^{y}\, \varphi(\boldsymbol{\eta}_{\mathbf{x}}), \tag{2.45}$$
$$\mathbf{x} \in \Omega_{h,\,x}^{ir} \cap \Omega_{h,\,y}^{ir}.$$

On note que le nombre d'inconnues est égal au nombre de points de l'ensemble Ω_h et qu'il s'agit d'un système algébrique linéaire de matrice carrée. On établit une estimation à priori pour en justifier la possibilité.

LEMME 2.3. *Soit u^h solution du système* (2.42) *à* (2.45), *où* $\varphi = 0$ *sur* Γ_h. *On a l'estimation*

$$\max_{\Omega_h} |u^h| \leqslant \frac{b^2(1+c_7)}{2c_7} \max_{\Omega_h^r} |f| + \frac{h^2}{2c_7} \max_{\Omega_h^{ir}} |f| . \qquad (2.46)$$

DÉMONSTRATION. Soit le système

$$-v^h_{x\bar{x}} - v^h_{y\bar{y}} = f \quad \text{sur} \quad \Omega_h^r ,$$
$$v^h = 0 \quad \text{sur} \quad \Omega_h^{ir} \cup \Gamma_h .$$

Le système vérifiant le principe du maximum admet une solution unique, et un théorème de comparaison a lieu (voir [41]). Ce théorème dit que si

$$-v_{x\bar{x}} - v_{y\bar{y}} = g \quad \text{sur} \quad \Omega_h^r ,$$
$$v \geqslant 0 \quad \text{sur} \quad \Omega_h^{ir} \cup \Gamma_h$$

alors l'inégalité

$$|f(\mathbf{x})| \leqslant g(\mathbf{x}) \quad \text{sur} \quad \Omega_h^r$$

implique

$$|v^h(\mathbf{x})| \leqslant v(\mathbf{x}) \quad \text{sur} \quad \Omega_h .$$

On pose

$$v(\mathbf{x}) = \frac{1}{4}(2b^2 - x^2 - y^2) \max_{\Omega_h^r} |f| .$$

Evidemment, $v(\mathbf{x}) \geqslant 0$ sur Ω_h tout entier, y compris sur $\Omega_h^{ir} \cup \Gamma_h$. En outre

$$-v_{x\bar{x}}(\mathbf{x}) - v_{y\bar{y}}(\mathbf{x}) = -\Delta v(\mathbf{x}) = \max_{\Omega_h^r} |f|$$

car il n'y a pas d'erreur d'approximation. Aussi

$$|f(\mathbf{x})| \leqslant g(\mathbf{x}) \qquad \text{sur} \quad \Omega_h^r ,$$
$$|v^h(\mathbf{x})| \leqslant v(\mathbf{x}) \leqslant \frac{b^2}{2} \max_{\Omega_h^r} |f| \quad \text{sur} \quad \Omega_h .$$

Le système

$$-w^h_{x\bar{x}} - w^h_{y\bar{y}} = 0 \quad \text{sur} \quad \Omega_h^r ,$$
$$w^h = u^h \qquad \text{sur} \quad \Omega_h^{ir} \cup \Gamma_h$$

possède également une solution unique. En identifiant w^h et la fonction constante

$$w(\mathbf{x}) = \max_{\Omega_h^{ir}} |u^h|$$

on trouve l'inégalité

$$|w^h(\mathbf{x})| \leqslant \max_{\Omega_h^{ir}} |u^h|$$

juste pour $\mathbf{x} \in \Omega_h$.

Quel que soit u^h fixé, la solution du système d'équations

$$-z_{\bar{x}x}^h - z_{\bar{y}y}^h = f \quad \text{sur} \quad \Omega_h^r,$$
$$z^h = u^h \quad \text{sur} \quad \Omega_h^{ir} \cup \Gamma_h$$

est unique, si bien que $z^h = u^h$ sur Ω_h^r car il vérifie lui aussi ce système. Vu que $z^h = v^h + w^h$, on a les estimations

$$|u^h(\mathbf{x})| = |z^h(\mathbf{x})| \leqslant |v^h(\mathbf{x})| + |w^h(\mathbf{x})| \leqslant$$
$$\leqslant \max_{\Omega_h^{ir}} |u^h| + \frac{b^2}{2} \max_{\Omega_h^r} |f| \qquad \forall \mathbf{x} \in \Omega_h^r. \tag{2.47}$$

Soit $\bar{\mathbf{x}}$ qui réalise

$$\max_{\Omega_h^{ir}} |u^h|.$$

Cela peut être, par exemple, le point

$$\bar{\mathbf{x}} = (\bar{x}, \bar{y}) \in \Omega_{h,x}^{ir} \cap \Omega_{h,y}^r.$$

On récrit (2.43) en recourant à la définition (2.40):

$$\left(\frac{2}{h^2} - \beta_{n,0}^x\right) u^h(\bar{x}, \bar{y}) - \frac{1}{h^2} u^h(\bar{x}, \bar{y} + h) - \frac{1}{h^2} u^h(\bar{x}, \bar{y} - h) -$$
$$-\sum_{i=1}^{np} \beta_{n,i}^x u^h(\bar{x} \pm ih, \bar{y}) = f(\bar{x}, \bar{y}). \tag{2.48}$$

Ici $n = l - 1$ et le signe de $\bar{x} \pm ih$ est sans importance. On a pour les nœuds $\mathbf{x} \in \Omega_h^{ir}$ l'inégalité

$$|u^h(\mathbf{x})| \leqslant |u^h(\bar{\mathbf{x}})|,$$

et (2.47) implique pour $\mathbf{x} \in \Omega_h^r$

$$|u^h(\mathbf{x})| \leqslant |u^h(\bar{\mathbf{x}})| + \frac{b^2}{2} \max_{\Omega_h^r} |f|. \tag{2.49}$$

On estime donc que (2.49) est vérifiée par chaque $\mathbf{x}$ de Ω_h. On se sert de cette inégalité pour évaluer les termes de (2.48) :

$$\left| \frac{2}{h^2} - \beta^h_{n,0} \right| \left| u^h(\bar{\mathbf{x}}) \right| \leqslant$$

$$\leqslant \left(\frac{2}{h^2} + \sum_{i=1}^{np} |\beta^x_{n,i}| \right) \left(|u^h(\bar{\mathbf{x}})| + \frac{b^2}{2} \max_{\Omega^r_h} |f| \right) + \max_{\Omega^{ir}_h} |f|. \quad (2.50)$$

L'inégalité (2.39) donne $\beta^x_{n,0} < 0$, si bien que le premier facteur dans (2.50) vaut

$$\frac{2}{h^2} + |\beta^x_{n,0}|.$$

Le premier facteur du second membre est positif. On divise par ce facteur et on effectue des transformations simples, il vient

$$\frac{\dfrac{2}{h^2} + |\beta^x_{n,0}|}{\dfrac{2}{h^2} + \sum_{i=1}^{np} |\beta^x_{n,i}|} |u^h(\bar{\mathbf{x}})| \leqslant |u^h(\bar{\mathbf{x}})| + \frac{b^2}{2} \max_{\Omega^r_h} |f| + \frac{h^2}{2} \max_{\Omega^{ir}_h} |f|.$$

On porte $|u^h(\bar{\mathbf{x}})|$ dans le premier membre et on utilise (2.39). On obtient

$$c_7 |u^h(\bar{\mathbf{x}})| \leqslant \frac{b^2}{2} \max_{\Omega^r_h} |f| + \frac{h^2}{2} \max_{\Omega^{ir}_h} |f|. \quad (2.51)$$

Si

$$\bar{\mathbf{x}} \in \Omega^r_{h,x} \cap \Omega^{ir}_{h,y},$$

des calculs analogues donnent (2.51). On divise membre à membre par c_7:

$$|u^h(\bar{\mathbf{x}})| \leqslant \frac{b^2}{2c_7} \max_{\Omega^r_h} |f| + \frac{h^2}{2c_7} \max_{\Omega^{ir}_h} |f|. \quad (2.52)$$

Le dernier cas à étudier est

$$\bar{\mathbf{x}} \in \Omega^{ir}_{h,x} \cap \Omega^{ir}_{h,y}.$$

Il lui correspond non la relation (2.48), mais l'égalité

$$(-\beta^x_{n,0} - \beta^y_{n,0})\, u^h(\bar{x}, \bar{y}) - \sum_{i=1}^{np} \beta^x_{n,i}\, u^h(\bar{x} \pm ih, \bar{y}) -$$

$$- \sum_{j=1}^{np} \beta^y_{n,j}\, u^h(\bar{x}, \bar{y} \pm jh) = f(\bar{x}, \bar{y}).$$

Ici $\beta^x_{n,i}$ sont les coefficients de l'opérateur J^x_n, $\beta^y_{n,r}$ ceux de J^y_n, et les signes dans $\bar{x} \pm ih$ et $\bar{y} \pm jh$ n'influent aucunement sur les calculs. Avec l'inégalité (2.49), on a

$$- \beta^x_{n,0} - \beta^y_{n,0} \mid \mid u^h(\bar{\mathbf{x}}) \mid \leqslant$$

$$\leqslant \left(\sum_{i=1}^{np} \mid \beta^x_{n,i} \mid + \sum_{j=1}^{np} \mid \beta^y_{n,j} \mid\right)\left(\mid u^h(\bar{\mathbf{x}}) \mid + \frac{b^2}{2} \max_{\Omega^r_h} \mid f \mid\right) + \max_{\Omega^{ir}_h} \mid f \mid. \quad (2.53)$$

La majoration (2.39) implique pour $\beta^x_{n,0}$ et $\beta^y_{n,0}$

$$\beta^x_{n,0} < 0, \qquad \beta^y_{n,0} < 0;$$

aussi le premier facteur dans le premier membre de (2.53) est égal à

$$\mid \beta^x_{n,0} \mid + \mid \beta^y_{n,0} \mid .$$

On ajoute au premier membre

$$\frac{4}{h^2} \mid u^h(\bar{\mathbf{x}}) \mid$$

et au second membre

$$\frac{4}{h^2}\left(\mid u^h(\bar{\mathbf{x}}) \mid + \frac{b^2}{2} \max_{\Omega^r_h} \mid f \mid\right),$$

et on divise par

$$\sum \mid \beta^x_{n,i} \mid + \sum \mid \beta^y_{n,j} \mid + \frac{4}{h^2},$$

il vient

$$\left(1 + \frac{\left|\beta^x_{n,0}\right| - \sum\limits_{i=1}^{np}\left|\beta^x_{n,i}\right| + \left|\beta^y_{n,0}\right| - \sum\limits_{j=1}^{np}\left|\beta^y_{n,j}\right|}{\frac{2}{h^2} + \sum\limits_{i=1}^{np}\left|\beta^x_{n,i}\right| + \frac{2}{h^2} + \sum\limits_{j=1}^{np}\left|\beta^y_{n,j}\right|}\right) \mid u^h(\bar{\mathbf{x}}) \mid \leqslant$$

$$\leqslant \mid u^h(\bar{\mathbf{x}}) \mid + \frac{b^2}{2} \max_{\Omega^r_h} \mid f \mid + \frac{h^2}{4} \max_{\Omega^{ir}_h} \mid f \mid . \quad (2.54)$$

Etant donné (2.39), les quantités

$$\alpha_1 = \mid \beta^x_{n,0} \mid - \sum \mid \beta^x_{n,i} \mid, \qquad \beta_1 = \frac{2}{h^2} + \sum \mid \beta^x_{n,i} \mid,$$

$$\alpha_2 = \mid \beta^y_{n,0} \mid - \sum \mid \beta^y_{n,j} \mid, \qquad \beta_2 = \frac{2}{h^2} + \sum \mid \beta^y_{n,j} \mid$$

vérifient les inégalités

$$\alpha_1/\beta_1 \geq c_7, \qquad \alpha_2/\beta_2 \geq c_7,$$

d'où

$$\frac{\alpha_1 + \alpha_2}{\beta_1 + \beta_2} \geq c_7. \tag{2.55}$$

La dernière estimation permet de déduire de (2.54)

$$c_7 \mid u \ (\bar{\mathbf{x}}) \leq \frac{b^2}{2} \max_{\Omega_h^r} |f| + \frac{h^2}{4} \max_{\Omega_h^{ir}} |f| .$$

On remplace le dernier terme par son majorant et on divise par c_7, ce qui redonne (2.52).

Les deux estimations (2.52) et (2.49) fournissent (2.46). Le lemme 2.3 se trouve démontré.

S'agissant du système homogène associé au problème aux différences (2.42) à (2.45), l'estimation (2.46) conduit de façon classique à l'existence et à l'unicité pour toute f et toute φ continues.

On construit la solution corrigée à partir de plusieurs solutions relatives à h différents. On pose $s = [l/2]$. On énonce pour chaque réseau $\Omega_{h/k}$ de pas h/k, $k = 1, \ldots, s+1$, le problème (2.42) à (2.45) et on en cherche la solution. Toutes les solutions $u^{h/k}$ sont définies sur Ω_h. On forme la combinaison linéaire

$$U^H(\mathbf{x}) = \sum_{k=1}^{s+1} \frac{2(-1)^{s-k+1} k^{2s+2}}{(s-k+1)!\,(s+k+1)!}\, u^{h/k}(\mathbf{x}), \qquad \mathbf{x} \in \Omega_h, \tag{2.56}$$

et on établit son ordre de précision.

THÉORÈME 2.4. *On suppose que les conditions du théorème* 1.2 *ont lieu pour le problème* (2.1), (2.2) *et que* $l \geq 2$, $\alpha \in (0, 1)$. *La solution corrigée* (2.56) *est évaluée par*

$$\max_{\Omega_h} | U^H - u | \leq c_9\, h^{l+\alpha}. \tag{2.57}$$

DÉMONSTRATION. On vérifie la validité des conditions du théorème 2.2, du chapitre premier sans oublier la remarque faite fin § 1.2. On pose

$$M_k(\Omega) = C^{k+\alpha}(\bar{\Omega}),$$
$$P_k(\bar{\Omega}) = C^{k+2+\alpha}(\bar{\Omega}),$$
$$N_k(D) = C^{k+2+\alpha}(\Gamma).$$

La condition A du § 1.2 découle alors du théorème 1.2 de ce chapitre. On pose, pour le problème discret (2.3) du § 1.2, $\bar{\Omega}_h = \Omega_h \cup \Gamma_h$, $\check{\Omega}_h = \Omega_h$, $D_h = \Gamma_h$, et on prend pour normes

$$\|u\|_{\bar{\Omega}_h} = \max_{\bar{\Omega}_h} |u|, \qquad \|u\|_{\check{\Omega}_h} = \max_{\Omega_h^r} |u| + h^2 \max_{\Omega_h^{ir}} |u|.$$

La condition B′ coïncide avec l'estimation à priori du lemme 2.3. Le développement de l'erreur d'approximation en série suivant les puissances paires de h résulte du procédé de construction du schéma aux différences, si bien que la condition D est vérifiée avec la constante $\beta = \alpha$, et la condition aux limites est approchée exactement, i.e. $l_h u = lu$ sur Γ_h. On utilise maintenant le théorème 3.2, § 1.3. La condition (3.15) est remplie avec la constante $d_3 = 1/s$. Ce théorème implique (2.57), c.q.f.d.

On note que le système (2.42) à (2.45) est à matrice non symétrique, ce qui restreint la classe correspondante de méthodes de résolution. On emploie cependant la technique décrite fin du no 4.2.1 qui ramène le système à une suite de problèmes algébriques de matrices symétriques. Le critère d'arrêt des procédés itératifs utilise largement l'inégalité (2.46) qui permet d'apprécier l'erreur sur une approximation de la solution du système (2.42) à (2.45) au vu du résidu.

4.3. Problème de Dirichlet dans un rectangle

Il est connu que la régularité de la solution est d'ordinaire détériorée au voisinage des points anguleux, ce qui empêche en général de raffiner les solutions approchées. La perte de régularité n'a cependant pas lieu si les entrées du problème vérifient certaines conditions de concordance, cas que nous allons considérer.

Soit l'équation

$$-\Delta u + au = f, \tag{3.1}$$

avec a non négative de $\Omega \subset \mathbf{R}^2$ qui est le rectangle

$$\Omega = \{(x_1, x_2);\ 0 < x_1 < b_1;\ 0 < x_2 < b_2\}.$$

On suppose les coefficients de (3.1) tels que

$$a,\ f \in C^{3+\alpha}(\bar{\Omega}), \qquad \alpha \in (0,\ 1). \tag{3.2}$$

On demande une solution satisfaisant à la condition aux limites homogène

$$u = 0 \quad \text{sur} \quad \Gamma, \tag{3.3}$$

Γ étant la frontière de Ω.

On note que le problème posé admet une solution dans $C^{5+\alpha}(\Omega)$, mais une régularité insuffisante de la frontière (en effet, $\Gamma \in C^{\gamma}$, $\gamma \in (0, 1)$) ne nous permet pas de parler d'une solution de classe $C^{5+\alpha}(\bar{\Omega})$, voire $C^{2+\alpha}(\bar{\Omega})$. Il suffit néanmoins d'exiger que le second membre remplisse certaines conditions de concordance pour que la solution appartienne nécessairement à des classes de régularité supérieure.

LEMME 3.1. *Soit, dans le problème* (3.1) *à* (3.3),

$$a, f \in C^{1+\alpha}(\bar{\Omega}), \qquad f(\mathbf{x}) = 0 \tag{3.4}$$

dans chaque angle du rectangle. Alors $u \in C^{3+\alpha}(\bar{\Omega})$.

DÉMONSTRATION. On a $\Gamma \in C^{\alpha}$, si bien qu'on conclut conformément à [26] à l'appartenance de la solution à $C^{\alpha}(\bar{\Omega})$. On porte le terme au dans le second membre et on pose le problème

$$\begin{aligned} -\Delta v &= f - au \quad \text{dans} \quad \Omega, \\ v &= 0 \qquad\quad\; \text{sur} \quad \Gamma. \end{aligned} \tag{3.5}$$

La fonction $f - au$ est dans $C^{\alpha}(\bar{\Omega})$, et il y a existence et unicité de la solution v. Aussi $v \equiv u$. Mais le problème (3.5) vérifie les conditions de concordance qui impliquent selon [8], [51] l'appartenance de la solution à $C^{2+\alpha}(\bar{\Omega})$: la différence $f(\mathbf{x}) - a(\mathbf{x})\, u(\mathbf{x})$ est nulle en chaque sommet du rectangle parce que $f(\mathbf{x})$ s'y annule en vertu de (3.4) et $u(\mathbf{x})$ l'est par suite de la condition aux limites (3.3). Aussi $u \in C^{2+\alpha}(\bar{\Omega})$, et, partant, le second membre de l'équation (3.5) est de classe $C^{1+\alpha}(\bar{\Omega})$. La condition de concordance étant remplie entraîne (voir [8], [51]) pour ce second membre la propriété de la solution d'être dans $C^{3+\alpha}(\bar{\Omega})$, i.e. on a le résultat du lemme.

Avec une deuxième condition de concordance, on garantit une régularité plus grande.

LEMME 3.2. *On suppose que le second membre de l'équation du problème* (3.1) *à* (3.3) *vérifie les conditions de concordance*

$$f(\mathbf{x}) = 0, \qquad \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) - \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) = 0 \tag{3.6}$$

dans chaque angle du rectangle. Alors $u \in C^{5+\alpha}(\bar{\Omega})$.

DÉMONSTRATION. Il est déjà apparu que $u \in C^{3+\alpha}(\bar{\Omega})$ (lemme 3.1). Soit le problème (3.5). Avec les hypothèses du lemme 3.2, le second membre est dans $C^{3+\alpha}(\bar{\Omega})$. Pour qu'il en découle $v \in C^{5+\alpha}(\bar{\Omega})$, il suffit d'exiger (voir [8], [51]) que la condition de concordance

$$\frac{\partial^2}{\partial x_1^2}(f - au) - \frac{\partial^2}{\partial x_2^2}(f - au) = 0$$

ait lieu dans chaque sommet du rectangle. Cette condition équivaut par suite de l'hypothèse du lemme à l'égalité

$$\frac{\partial^2}{\partial x_1^2}(au) - \frac{\partial^2}{\partial x_2^2}(au) = 0.$$

Comme les dérivées secondes de au sont continues sur $\bar{\Omega}$, on les calcule sur les côtés du rectangle. Or $au = 0$ sur Γ en vertu de la condition aux limites. Aussi la condition de concordance pour le problème (3.5) coïncide avec l'hypothèse du lemme et $v \in C^{5+\alpha}(\bar{\Omega})$. Du moment que $v = u$, le lemme se trouve démontré.

On construit le schéma aux différences. Soit le réseau rectangulaire régulier

$$\bar{\Omega}_h = \{(x_1, x_2);\ x_1 = ih_1,\ x_2 = jh_2,\ 0 \leqslant i,\ j \leqslant N\}$$

de pas $h_1 = b_1/N$ et $h_2 = b_2/N$. On introduit les notations

$$\Gamma_h = \bar{\Omega}_h \cap \Gamma, \qquad \Omega_h = \bar{\Omega}_h \setminus \Gamma_h.$$

On discrétise le problème (3.1) à (3.3) par le schéma usuel à cinq points

$$-u^h_{\bar{x}_1 x_1}(\mathbf{x}) - u^h_{\bar{x}_2 x_2}(\mathbf{x}) + a(\mathbf{x})\, u^h(\mathbf{x}) = f(\mathbf{x}), \qquad \mathbf{x} \in \Omega_h, \tag{3.7}$$

avec les conditions aux limites

$$u^h(\mathbf{x}) = 0, \qquad \mathbf{x} \in \Gamma_h. \tag{3.8}$$

LEMME 3.3. *On a pour le problème aux différences* (3.7), (3.8) *l'inégalité*

$$\max_{\mathbf{x} \in \bar{\Omega}_h} |u^h(\mathbf{x})| \leqslant \frac{b_1^2 + b_2^2}{16} \max_{\mathbf{x} \in \bar{\Omega}} |f(\mathbf{x})|. \tag{3.9}$$

DÉMONSTRATION. Un théorème de comparaison démontré dans [42] permet de dire que la solution du problème

$$-W_{\bar{x}_1 x_1}(\mathbf{x}) - W_{\bar{x}_2 x_2}(\mathbf{x}) + a(\mathbf{x})\, W(\mathbf{x}) = g_1(\mathbf{x}), \qquad \mathbf{x} \in \Omega_h,$$
$$W(\mathbf{x}) = g_2(\mathbf{x}), \qquad \mathbf{x} \in \Gamma_h,$$

avec les conditions

$$|f(\mathbf{x})| \leqslant g_1(\mathbf{x}) \quad \text{sur} \quad \Omega_h, \qquad 0 \leqslant g_2(\mathbf{x}) \quad \text{sur} \quad \Gamma_h$$

majore la solution du problème (3.7), (3.8):

$$|u^h(\mathbf{x})| \leqslant W(\mathbf{x}) \quad \text{sur} \quad \bar{\Omega}_h.$$

On prend pour W la fonction majorante de Gerschgorin

$$W(\mathbf{x}) = \frac{1}{4}\left(\frac{b_1^2 + b_2^2}{4} - \left(x_1 - \frac{b_1}{2}\right)^2 - \left(x_2 - \frac{b_2}{2}\right)^2\right) \max_{\Omega_h} |f(\mathbf{x})|.$$

On établit sans peine que $W(\mathbf{x}) \geq 0$ sur le rectangle $\bar{\Omega}_h$ tout entier. La fonction $W(\mathbf{x})$ a de plus, à partir des dérivées troisièmes, toutes ses dérivées nulles. Aussi les dérivées secondes aux différences sont exactement les dérivées partielles correspondantes:

$$W_{\bar{x}_i x_i}(\mathbf{x}) = \frac{\partial^2 W}{\partial x_i^2}(\mathbf{x}), \qquad \mathbf{x} \in \Omega_h.$$

Etant donné ce résultat, on a

$$g_1(\mathbf{x}) = -\Delta W(\mathbf{x}) + a(\mathbf{x}) W(\mathbf{x}) \geq -\Delta W(\mathbf{x}) = \\ = \max_{\Omega_h} |f| \geq |f(\mathbf{x})|.$$

Ainsi, les conditions du théorème de comparaison sont remplies et

$$u^h(\mathbf{x}) \leq W(\mathbf{x}), \qquad \mathbf{x} \in \bar{\Omega}_h.$$

D'où

$$\max_{\bar{\Omega}_h} |u^h(\mathbf{x})| \leq \max_{\bar{\Omega}_h} |W(\mathbf{x})| \leq \frac{b_1^2 + b_2^2}{16} \max_{\Omega_h} |f(\mathbf{x})|.$$

Le lemme est démontré.

On pose $h = 1/N$ et on cherche le développement de la solution approchée suivant les puissances de h.

THÉORÈME 3.4. *On suppose qu'on est pour le problème* (3.1) *à* (3.3) *dans les hypothèses du lemme* 3.2. *Alors la solution* u^h *du problème discret* (3.7), (3.8) *admet le développement*

$$u^h = u + h^2 v + h^{3+\alpha} \eta^h \quad \text{sur} \quad \bar{\Omega}_h \tag{3.10}$$

où v *est une fonction continue indépendante de* h *et la fonction discrète* η^h *est bornée*:

$$\max_{\bar{\Omega}_h} |\eta^h| \leq c_1. \tag{3.11}$$

DÉMONSTRATION. On cherche v à partir de la solution du problème différentiel

$$-\Delta v + av = \frac{b_1^2}{12} \frac{\partial^4 u}{\partial x_1^4} + \frac{b_2^2}{12} \frac{\partial^4 u}{\partial x_2^4} \quad \text{dans} \quad \Omega, \tag{3.12}$$

$$v = 0 \quad \text{sur} \quad \Gamma.$$

On vérifie d'abord que $v \in C^{3+\alpha}(\bar{\Omega})$. En effet, si l'on est dans les conditions du lemme 3.2, alors $u \in C^{5+\alpha}(\bar{\Omega})$, si bien que le second membre de (3.12) appartient à $C^{1+\alpha}(\bar{\Omega})$. On est donc, pour le problème (3.12), dans l'hypothèse de régularité du second membre du

lemme 3.1. La condition de concordance est satisfaite parce que $\partial^4 u/\partial x_1^4$ et $\partial^4 u/\partial x_2^4$ s'annulent dans chaque angle du rectangle (on l'établit assez facilement du moment que les dérivées étant continues sur $\bar{\Omega}$ sont calculées le long des côtés du rectangle sur lesquels $u = 0$).

Ainsi, les fonctions u^h, u et v sont définies de façon unique. On pose

$$\eta^h = (u^h - u - h^2 v)\, h^{-3-\alpha} \quad \text{sur} \quad \bar{\Omega}_h \tag{3.13}$$

et on la porte dans l'opérateur aux différences du problème (3.7):

$$-\eta^h_{\bar{x}_1 x_1} - \eta^h_{\bar{x}_2 x_2} + a\eta^h = \{-u^h_{\bar{x}_1 x_1} - u^h_{\bar{x}_2 x_2} + au^h + u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2} - au + h^2 (v_{\bar{x}_1 x_1} + v_{\bar{x}_2 x_2} - av)\}\, h^{-3-\alpha}.$$

On transforme cette expression à l'aide de l'équation (3.7) et du développement (2.4) des dérivées discrètes, il vient

$$-\eta^h_{\bar{x}_1 x_1} - \eta^h_{\bar{x}_2 x_2} + a\eta^h = \left(f + \frac{\partial^2 u}{\partial x_1^2} + \frac{h_1^2}{12} \frac{\partial^4 u}{\partial x_1^4} + \frac{\partial^2 u}{\partial x_2^2} + \frac{h_2^2}{12} \frac{\partial^4 u}{\partial x_2^4} - au - h^2 av + h^2 \frac{\partial^2 v}{\partial x_1^2} + h^2 \frac{\partial^2 v}{\partial x_2^2} + h_1^{3+\alpha} \xi_1 + h_2^{3+\alpha} \xi_2 \right) h^{-3-\alpha}.$$

Ici $|\xi_1(\mathbf{x})| \leqslant c_2$, $|\xi_2(\mathbf{x})| \leqslant c_3$ uniformément par rapport à $\mathbf{x}$ et h. On rappelle que $h_1 = b_1 h$, $h_2 = b_2 h$. On utilise l'équation (3.1):

$$-\eta^h_{\bar{x}_1 x_1} - \eta^h_{\bar{x}_2 x_2} + a\eta^h = \left(\frac{b_1^2}{12} \frac{\partial^4 u}{\partial x_1^4} + \frac{b_2^2}{12} \frac{\partial^4 u}{\partial x_2^4} + \Delta v - av \right) h^{-1-\alpha} + \xi_3,$$

avec

$$|\xi_3(\mathbf{x})| \leqslant c_4 = c_2 b_1^{3+\alpha} + c_3 b_2^{3+\alpha}. \tag{3.14}$$

Le terme en $h^{-1-\alpha}$ disparaît par suite de la définition de v, ce qui donne

$$-\eta^h_{\bar{x}_1 x_1} - \eta^h_{\bar{x}_2 x_2} + a\eta^h = \xi_3 \quad \text{sur} \quad \bar{\Omega}_h. \tag{3.15}$$

L'égalité (3.13) sur Γ_h entraîne de plus

$$\eta^h = 0 \quad \text{sur} \quad \Gamma_h \tag{3.16}$$

du moment que les trois fonctions u^h, u, v s'annulent sur Γ_h. On a pour le problème (3.15), (3.16) l'estimation du lmmee 3.3, d'où

$$\max_{\bar{\Omega}_h} |\eta^h(\mathbf{x})| \leqslant c_4 \frac{b_1^2 + b_2^2}{16},$$

avec c_4 définie dans (3.14). On vient donc d'établir l'inégalité (3.11), ce qui achève la démonstration du théorème.

Appliquons le théorème à un cas concret. On suppose résolus deux problèmes (3.7), (3.8) pour les pas h et $h/2$. Le réseau Ω_h est le domaine de définition de deux solutions approchées u^h et $u^{h/2}$. On les additionne avec les poids $-1/3$ et $4/3$ et on obtient une solution approchée dont la précision est $O(h^{3+\alpha})$:

$$\max_{\Omega_h} \left| u(\mathbf{x}) - \left(\frac{4}{3} u^{h/2}(\mathbf{x}) - \frac{1}{3} u^h(\mathbf{x}) \right) \right| \leqslant h^{3+\alpha} c_5. \qquad (3.17)$$

la constante c_5 étant indépendante de h et $\mathbf{x}$. La démonstration s'appuie sur le développement (3.10). Son idée est la même que dans les paragraphes précédents: elle consiste à choisir les poids égaux à $4/3$ et $-1/3$ pour faire disparaître les termes en h^2 de la combinaison linéaire.

Les résultats obtenus demandent des remarques.

REMARQUE 1. Il est clair que le reste augmente avec la perte de régularité du second membre, disons pour $f \in C^{2+\alpha}(\bar{\Omega})$ et que le développement s'écrit

$$u^h = u + h^2 v + h^{2+\alpha} \eta^h \quad \text{sur} \quad \bar{\Omega}_h.$$

D'autre part, on n'arrive pas à prolonger sensiblement le développement (à diminuer la grandeur du reste irrégulier) en élevant la régularité ($f \in C^{k+\alpha}(\bar{\Omega})$, $k \geqslant 4$) même si l'on fait l'hypothèse de solution aussi régulière qu'on le veut. Le fait est que les problèmes auxiliaires relatifs à la recherche des fonctions v_i du développement

$$u^h = u + h^2 v_1 + h^4 v_2 + \ldots + h^{r+\alpha} \eta^h \qquad (3.18)$$

ne vérifient pas nécessairement les conditions de concordance qui garantissent la régularité de la solution. Or ces conditions sont non seulement suffisantes, mais encore nécessaires pour l'existence d'une solution régulière (voir [8], [51]).

On propose dans [9] de remplacer la solution u par la somme $W + Q$, où W est la fonction inconnue vérifiant la même équation avec un autre second membre, et Q une fonction connue. On a prouvé qu'on obtient finalement une équation pour W suffisamment régulière et que les conditions de concordance sont vérifiées, à part W, par tous les problèmes auxiliaires, ce qui garantit la régularité de leurs solutions. Ainsi, on a établi dans [9] le développement (3.18) sous l'hypothèse $f \in C^{r+\alpha}(\bar{\Omega})$.

REMARQUE 2. Le développement (3.10) aidant, on justifie la recherche des dérivées premières et secondes de la solution par recours à la solution discrète. Par exemple,

$$\frac{u^h(x_1+h_1,\ x_2)-u^h(x_1-h_1,\ x_2)}{2h_1}=\frac{\partial u}{\partial x}(x_1,\ x_2)+h^2\xi_3(x_1,\ x_2),$$

avec $|\xi_3(\mathbf{x})|\leqslant c_7$ indépendant de $\mathbf{x}$ et h. En effet, on a moyennant (3.10)

$$\frac{u(x_1+h_1,\ x_2)-u(x_1-h_1,\ x_2)}{2h_1}+h^2\frac{v(x_1+h_1,\ x_2)-v(x_1-h,\ x_2)}{2h_1}+$$
$$+h^{3+\alpha}\frac{\eta^h(x_1+h_1,\ x_2)-\eta^h(x_1-h_1,\ x_2)}{2h_1}=\frac{\partial u}{\partial x_1}(x_1,\ x_2)+$$
$$+\frac{h_1^2}{3}\frac{\partial^3 u}{\partial x_1^3}(\theta_1,\ x_2)+h_1^2\frac{\partial v}{\partial x_1}(\theta_2,\ x_2)+$$
$$+\frac{h^{2+\alpha}}{2b_1}(\eta^h(x_1+h_1,\ x_2)-\eta^h(x_1-h_1,\ x_2)),$$
$$\theta_i\in(x_1-h_1,\ x_1+h_1).$$

On évalue les derniers termes et on a l'estimation du reste:

$$|\xi_3(\mathbf{x})|\leqslant\frac{b_1^2}{3}\max_{\bar\Omega}\left|\frac{\partial^3 u}{\partial x_1^3}\right|+b_1^2\max_{\bar\Omega}\left|\frac{\partial v}{\partial x_1}\right|+\frac{c_1}{b_1}.$$

On procède de même en ce qui concerne $\partial u/\partial y$. Les dérivées secondes sont calculées avec la précision $O(h^{1+\alpha})$. Par exemple,

$$\frac{\partial^2 u}{\partial x_1^2}(\mathbf{x})=u^h_{\bar x_1 x_1}(\mathbf{x})+h^{1+\alpha}\xi_4(\mathbf{x}).$$

En effet, on utilise (3.10), il vient

$$u^h_{\bar x_1 x_1}(\mathbf{x})=u_{\bar x_1 x_1}(\mathbf{x})+h^2v_{\bar x_1 x_1}(\mathbf{x})+h^{3+\alpha}\eta^h_{\bar x_1 x_1}(\mathbf{x})=$$
$$=\frac{\partial^2 u}{\partial x_1^2}(\mathbf{x})+\frac{h_1^2}{12}\frac{\partial^4 u}{\partial x_1^4}(\theta_3,\ x_2)+h^2\frac{\partial^2 v}{\partial x_1^2}(\theta_4,\ x_2)+$$
$$+\frac{h^{1+\alpha}}{b_1^2}(\eta^h(x_1+h_1,\ x_2)-2\eta^h(x_1,\ x_2)+\eta^h(x_1-h_1,\ x_2)),$$

avec $\theta_i\in(x_1-h_1,\ x_1+h_1)$. On évalue les termes du second membre :

$$|\xi_4(\mathbf{x})|\leqslant\frac{b_1^2}{12}\max_{\Omega}\left|\frac{\partial^4 u}{\partial x_1^4}\right|+\max_{\Omega}\left|\frac{\partial^2 v}{\partial x_1^2}\right|+4\frac{c_2}{b_1^2}.$$

On construit une formule analogue pour $\partial^2 u/\partial x_2^2$. Quant à la dérivée partielle mixte, on l'obtient par différences centrales avec la même précision :

$$\frac{\partial^2 u}{\partial x_1 \partial x_2}(x_1, x_2) = \left(u^h(x_1 + h_1, x_2 + h_2) - u^h(x_1 - h_1, x_2 + h_2) + \right.$$
$$\left. + u_h(x_1 - h_1, x_2 - h_2) - u^h(x_1 + h_1, x_2 - h_2)\right) h_1^{-1} h_2^{-1} + O(h^{1+\alpha}).$$

4.4. Equation quasi linéaire dans un domaine triangulaire

S'agissant des équations quasi linéaires du type elliptique pour des problèmes en dimension deux dans un domaine avec points anguleux, on constate de règle que les singularités apparaissant au voisinage de ces points altèrent la régularité de la solution. Tout domaine polygonal étant un ensemble de domaines triangulaires, c'est aux sommets de ces derniers qu'on observe les singularités les plus caractéristiques. Les auteurs se proposent d'élaborer, à la lumière d'un problème concret, un algorithme qui garantit pour des entrées non concordantes une solution dont l'ordre de précision en h est voisin de deux. On fait ainsi le premier pas vers une solution notablement améliorée.

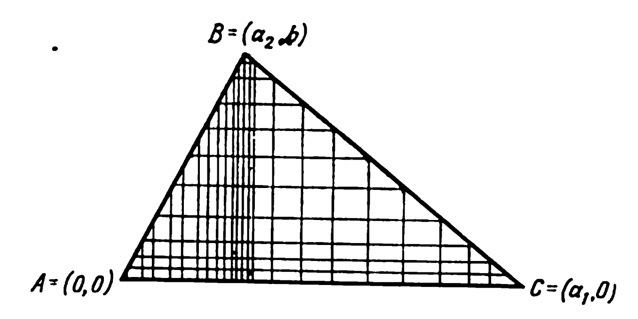


Fig. 4.6. Triangle couvert d'un réseau irrégulier

Dans ce paragraphe, on examinera l'influence des points anguleux qu'on étudiera en resserrant le réseau dans le cas d'une équation quasi linéaire et d'un domaine triangulaire pour un schéma discret usuel à cinq points.

Soit Ω un triangle ouvert de sommets $(0, 0)$, $(a_1, 0)$, (a_2, b), avec $b > 0$, $a_1 > a_2 > 0$ (fig. 4.6).

On considère l'équation

$$-\Delta u(\mathbf{x}) = f(u(\mathbf{x}), \mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{4.1}$$

où

$$f(p, \mathbf{x}) \in C^2((-\infty, \infty) \times \bar{\Omega}). \tag{4.2}$$

On cherche une solution qui vérifie la condition homogène

$$u = 0 \quad \text{sur} \quad \Gamma. \tag{4.3}$$

Γ étant la frontière de Ω. On admet que le problème possède au moins une solution de classe $L_2(\Omega)$ bornée en valeur absolue, i.e.

$$\operatorname{vrai} \sup_{\Omega} |u| \leqslant c_0. \tag{4.4}$$

On suppose de plus qu'on est dans la condition

$$\frac{\partial f}{\partial p}(p, \mathbf{x}) \leqslant 0, \qquad p \in (-\infty, \infty), \qquad \mathbf{x} \in \bar{\Omega}. \tag{4.5}$$

LEMME 4.1. *Soit $u \in L_2(\Omega)$ une solution qui remplit la condition (4.4). On suppose que le second membre f vérifie (4.2) et (4.5). Alors la solution u est unique et appartient à*

$$W_2^2(\Omega) \cap C^{\gamma}(\bar{\Omega}) \qquad \forall \gamma \in (0, 1).$$

DÉMONSTRATION. La condition (4.4) fait que $f(u(\mathbf{x}), \mathbf{x}) \in L_2(\Omega)$ en tant que fonction de $\mathbf{x}$. Aussi la solution du problème

$$-\Delta v = f(u(\mathbf{x}), \mathbf{x}) \quad \text{dans} \quad \Omega,$$
$$v = 0 \quad \text{sur} \quad \Gamma$$

est dans $W_2^2(\Omega)$ par suite du théorème 1.4. Comme la fonction u satisfait à cette équation, on a $v = u$ et, partant, $u \in W_2^2(\Omega)$. D'après un théorème de l'immersion (voir [26]), $u \in C^{\gamma}(\bar{\Omega})$ $\forall \gamma \in (0, 1)$. On montre que le problème (4.1), (4.3) ne possède pas d'autre solution vérifiant (4.4). En effet, soit u_1 une deuxième solution ayant cette propriété. Alors $u_1 \in W_2^2(\Omega)$, et la différence $w = u_1 - u$ vérifie l'identité

$$\int_{\Omega} [-\Delta w + f(u, \mathbf{x}) - f(u_1, \mathbf{x})]\, w \, d\mathbf{x} = 0$$

qui découle de (4.1). On utilise les conditions aux limites homogènes et la première formule de Green, il vient l'égalité

$$\int_{\Omega} \left[\left(\frac{\partial w}{\partial x_1} \right)^2 + \left(\frac{\partial w}{\partial x_2} \right)^2 + (f(u, \mathbf{x}) - f(u_1, \mathbf{x}))\, w \right] d\mathbf{x} = 0.$$

On applique la formule de Lagrange au dernier groupe de termes:

$$f(u, \mathbf{x}) - f(u_1, \mathbf{x}) = \frac{\partial f}{\partial p}(\xi, \mathbf{x})\,(u - u_1),$$

si bien qu'on obtient moyennant (4.5)

$$\int_{\Omega} \left[\left(\frac{\partial w}{\partial x_1} \right)^2 + \left(\frac{\partial w}{\partial x_2} \right)^2 \right] d\mathbf{x} \leqslant 0.$$

D'où $w = 0$ dans l'espace $\overset{\circ}{W}{}_2^1(\Omega)$. Aux termes du théorème mentionné, $u = u_1$ dans $L_2(\Omega)$. Le lemme est démontré.

Quant à l'existence d'une solution de $L_2(\Omega)$ du problème (4.1), (4.3), on trouve dans la monographie [26] plusieurs critères suffisants pour de nombreux problèmes quasi linéaires.

On note que la condition (4.5) peut être affaiblie (ce qui complique quelque peu les calculs suivants). On remplace à cet effet le zéro du second membre par un nombre positif $d_1 < 1/d_2^2$, avec d_2 la constante de l'inégalité de Friedrichs (voir [34])

$$\|u\|_{L_2(\Omega)} \leqslant d_2 |u| \qquad \forall u \in \overset{\circ}{W}{}_2^1(\Omega).$$

Les classes de régularité $C^{l+\alpha}(\bar{\Omega})$ utilisées plus haut ne donnent pas une idée complète de la solution du problème (4.1), (4.3) car si l'on n'est pas dans certaines conditions auxiliaires, la solution peut ne pas appartenir à $C^2(\bar{\Omega})$ même pour des seconds membres analytiques. Ce phénomène nous est familier dès le paragraphe précédent où l'on a réussi à énoncer les conditions de concordance sous forme explicite. Le cas présent est plus délicat. En effet, si les angles du triangle ne valent pas $\pi/k(k \geqslant 2$ étant un entier), alors les conditions déterminant la régularité de la solution se compliquent au point de devenir des relations intégrales de vérification pratique difficile.

D'autre part, les points anguleux n'influent pas sur la régularité à l'intérieur du domaine. Soit le problème

$$\begin{aligned} -\Delta v(\mathbf{x}) &= f(u(\mathbf{x}), \mathbf{x}), & \mathbf{x} &\in \Omega, \\ v(\mathbf{x}) &= 0, & \mathbf{x} &\in \Gamma. \end{aligned} \tag{4.6}$$

La fonction $f(u(\mathbf{x}), \mathbf{x})$ est dans $C^\gamma(\bar{\Omega})$, et $v \in C^{2+\gamma}(\Omega)$ par le théorème 1.1. Comme il y a unicité pour les problèmes (4.6) et (4.1), (4.3), on a $v = u$, donc $u \in C^{2+\gamma}(\Omega)$.

Le fait d'être de classes $C^{l+\gamma}(\Omega)$ n'impose cependant aucune restriction sur la grandeur des dérivées et de la fonction même (qui peuvent être non bornées sur Ω). Aussi la construction du schéma aux différences doit être précédée d'une étude du comportement de ces fonctions.

On suit les auteurs [8], [18], [38], [51]. On fixe un point anguleux et on introduit dans son voisinage distant d'une quantité positive des sommets restants les coordonnées polaires (r, φ) de centre le sommet concerné. On se propose, par exemple, de préciser le comportement de la solution u dans une région autour de $(0, 0)$. Cette région est prise égale au secteur circulaire $S = \{(r, \varphi);\ 0 < r < \varepsilon,\ 0 < \varphi < \beta\}$. Les coordonnées cartésiennes sont liées aux coordonnées polaires par les égalités $x_1 = r \sin \varphi$, $x_2 = r \cos \varphi$; β est la grandeur de l'angle BAC et ε est la demi-longueur du plus petit côté du trian-

gle. Dans ce cas, $S \subset \Omega$ et deux rayons de S sont confondus avec deux côtés du triangle. Conformément à [18], la solution v du problème (4.6) admet la représentation

$$v = w + \sum_{i=1}^{M} r^{\mu_i} \ln^{q_i} r \theta_i\ (\varphi). \tag{4.7}$$

La composante régulière * w est dans $C^{2+\gamma}(\bar{\Omega})$, $\{\mu_i\}_{i=1}^{M}$ est une suite non décroissante de nombres positifs, $\{q_i\}_{i=1}^{M}$ une famille de nombres entiers non négatifs et $\{\theta_i\ (\varphi)\}_{i=1}^{M}$ une famille de polynômes trigonométriques ∞-dérivables. Seul nous intéresse le terme principal du développement asymptotique (4.7), qui déterminera dans la suite le degré de resserrement du réseau. Son comportement est mis en lumière par plusieurs procédés élaborés en théorie des équations aux dérivées partielles, mais l'étude approfondie de ces procédés n'entre pas dans nos intentions. En coordonnées polaires, l'équation

$$-\Delta v = g \quad \text{dans} \quad S,$$

avec $g(\mathbf{x}) = f\big(u\ (\mathbf{x}),\ \mathbf{x}\big)$, s'écrit

$$-\frac{\partial^2 v}{\partial r^2} - \frac{1}{r}\frac{\partial v}{\partial r} - \frac{1}{r^2}\frac{\partial^2 v}{\partial \varphi^2} = g(r,\ \varphi) \quad \text{dans} \quad S. \tag{4.8}$$

Cela étant, on a

$$\begin{aligned} &v(r,\ 0) = 0, && r \in (0,\ \varepsilon), \\ &v(r,\ \beta) = 0, && r \in (0,\ \varepsilon), \\ &v(\varepsilon,\ \varphi) = u(\varepsilon,\ \varphi), && \varphi \in [0,\ \beta]. \end{aligned} \tag{4.9}$$

Il y a existence et unicité pour $r = 0$ si $|v\ (\delta,\ \varphi)| < \infty$ pour δ positifs tendant vers 0. On note que la distance entre l'arc $(\varepsilon,\ \varphi)$ et les sommets du triangle est positive, si bien que $u(\varepsilon, \varphi) \in C^{2+\gamma}[0,\ \beta]$ en tant que fonction de la variable indépendante φ.

On développe v et g en série de Fourier:

$$\begin{aligned} v\ (r,\ \varphi) &= \sum_{k=1}^{\infty} v_k\ (r) \sin k\lambda\varphi, \\ g\ (r,\ \varphi) &= \sum_{k=1}^{\infty} g_k\ (r) \sin k\lambda\varphi, \end{aligned} \tag{4.10}$$

* Plus précisément, [18] entraîne $w \in W_2^4\ (\Omega)$, mais $w \in C^{2+\gamma}\ (\bar{\Omega})\ \forall \gamma \in (0,1)$ aux termes d'un théorème de l'immersion (voir [26]).

où $\lambda = \pi/\beta$. On calcule les coefficients de Fourier par les formules

$$\begin{aligned} v_k(r) &= \frac{2}{\beta}\int_0^\beta v(r, \varphi) \sin k\lambda\varphi \, d\varphi, \\ g_k(r) &= \frac{2}{\beta}\int_0^\beta g(r, \varphi) \sin k\lambda\varphi \, d\varphi. \end{aligned} \tag{4.11}$$

On porte les développements (4.10) dans (4.8) et on identifie les coefficients des fonctions sin $k\lambda\varphi$, il vient

$$-v_k'' - \frac{1}{r} v_k' + \frac{k^2 \lambda^2}{r^2} v_k = g_k, \qquad k = 1, 2, \ldots$$

Les égalités (4.9), (4.11) fournissent les conditions

$$v_k(\varepsilon) = z_k \equiv \frac{2}{\beta}\int_0^\beta u(\varepsilon, \varphi) \sin k\lambda\varphi \, d\varphi,$$

$|v_k(\delta)| < \infty$ pour δ tendant vers zéro par valeurs positives.

Il résulte de [38] que le problème posé possède une solution

$$v_k(r) = -\frac{r^{k\lambda}}{2k\lambda\varepsilon^{2k\lambda}}\int_0^\varepsilon g_k(\rho)\rho^{k\lambda+1}\, d\rho + \frac{r^{k\lambda}}{2k\lambda}\int_r^\varepsilon g_k(\rho)\rho^{1-k\lambda}\, d\rho + \\ + \frac{r^{-k\lambda}}{2k\lambda}\int_0^r g_k(\rho)\rho^{k\lambda+1}\, d\rho + \frac{r^{k\lambda}}{\varepsilon^{k\lambda}} z_k$$

et une seule. On établit par des raisonnements simples mais fastidieux que c'est le cœfficient $v_1(r)$ qui est le moins régulier. Deux cas peuvent se présenter. Si $\beta = \pi/2$, alors $\lambda = 2$, et le terme principal du développement asymptotique (4.7) est un terme en $r^2 \ln r$. Dans le cas contraire, c'est des termes en r^λ. Aussi $v_1(r) \sin \lambda\varphi \in C^1(\overline{S})$. Cela est vrai à fortiori des autres termes de la série de Fourier de la fonction v, et on démontre la convergence absolue des séries de Fourier de $\partial v/\partial x$ et $\partial v/\partial y$. Donc $v \in C^1(\overline{S})$. Il y a plus. On a

$$r^\gamma \frac{\partial v_k}{\partial x}(r) \sin k\lambda\varphi \in C^\gamma(\overline{S}), \qquad k = 1, 2, \ldots, \gamma \in (0, 1),$$

d'où $r^\gamma \dfrac{\partial v}{\partial x} \in C^\gamma(\overline{S})$. De même, $r^\gamma \dfrac{\partial v}{\partial y} \in C^\gamma(\overline{S})$.

On introduit une fonction $d(\mathbf{x})$ égale à la distance de $\mathbf{x}$ et du plus proche sommet du triangle. Il est clair que $d(\mathbf{x}) = r$ à l'intérieur de

S. Aussi $d^\gamma \frac{\partial v}{\partial x}$, $d^\gamma \frac{\partial v}{\partial y} \in C^\gamma(\bar{S})$. Comme ces raisonnements sont justes pour tout sommet du triangle, on a

$$u \in C^1(\bar{\Omega}), \qquad d^\gamma \frac{\partial u}{\partial x}, \qquad d^\gamma \frac{\partial u}{\partial y} \in C^\gamma(\bar{\Omega}). \tag{4.12}$$

Ces propriétés donnent une information supplémentaire sur le second membre g de l'équation (4.8). On tire de (4.12) et des propriétés de f

$$g \in C^1(\bar{\Omega}), \qquad d^\gamma \frac{\partial g}{\partial x}, \qquad d^\gamma \frac{\partial g}{\partial y} \in C^\gamma(\bar{\Omega}).$$

On a par des considérations analogues pour les dérivées secondes et troisièmes de u:

$$\max_{\bar{\Omega}} \left(\left| d \frac{\partial^2 u}{\partial x^2} \right|, \left| d \frac{\partial^2 u}{\partial y^2} \right|, \left| d^2 \frac{\partial^3 u}{\partial x^3} \right|, \left| d^2 \frac{\partial^3 u}{\partial y^3} \right| \right) \leqslant c_1,$$

$$\max_{\mathbf{x}, \mathbf{x}' \in \bar{\Omega}} \left(\frac{\left| \frac{\partial^3 u}{\partial x^3}(\mathbf{x}) - \frac{\partial^3 u}{\partial x^3}(\mathbf{x}') \right|}{|(\mathbf{x} - \mathbf{x}')|^\alpha} \min \{ d^{2+\alpha}(\mathbf{x}), d^{2+\alpha}(\mathbf{x}') \} \right) \leqslant c_1, \tag{4.13}$$

$$\max_{\mathbf{x}, \mathbf{x}' \in \bar{\Omega}} \left(\frac{\left| \frac{\partial^3 u}{\partial y^3}(\mathbf{x}) - \frac{\partial^3 u}{\partial y^3}(\mathbf{x}') \right|}{|(\mathbf{x} - \mathbf{x}')|^\alpha} \min \{ d^{2+\alpha}(\mathbf{x}), d^{2+\alpha}(\mathbf{x}') \} \right) \leqslant c_1,$$

où $\alpha \in (0, 1)$ est quelconque, mais la constante c_1 dépend de α. Dans la suite, on fixe donc $\alpha \in (0, 1)$. En plus qu'on a les conditions (4.13), le théorème 1.1 entraîne $u \in C^{3+\alpha}(\Omega)$.

On passe à la discrétisation. On introduit la fonction

$$\varphi(t) = bt^2(3 - 2t)$$

et on note que $\varphi(0) = 0$, $\varphi(1) = b$. On fixe $N \geqslant 2$ entier et on pose $h = 1/N$. On munit le domaine d'une famille de droites parallèles

$$x_2 = \varphi(ih), \qquad i = 1, \ldots, N - 1,$$

et on mène par les points d'intersection des droites et des côtés AB et BC une deuxième famille de droites parallèles entre elles et orthogonales aux premières (voir fig. 4.6). On a donc couvert le triangle d'un réseau de discrétisation compatible avec sa frontière. On désigne par Ω_h l'ensemble des points de Ω en lesquels les droites de la première famille rencontrent celles de la seconde et par Γ_h l'ensemble des points d'intersection des droites et de la frontière du triangle. On pose $\bar{\Omega}_h = \Omega_h \cup \Gamma_h$.

On résout le problème (4.1), (4.3) par le schéma usuel à cinq points associé au réseau rectangulaire irrégulier:

$$L^h u^h(\mathbf{x}) \equiv - u^h_{\bar{x}_1 \hat{x}_1}(\mathbf{x}) - u^h_{\bar{x}_2 \hat{x}_2}(\mathbf{x}) = f(u^h(\mathbf{x}), \mathbf{x}), \qquad \mathbf{x} \in \Omega_h. \tag{4.14}$$

avec les conditions aux limites

$$u^h(\mathbf{x}) = 0, \qquad \mathbf{x} \in \Gamma_h. \tag{4.15}$$

La notation symbolique des dérivées aux différences traduit la propriété suivante: si l'on a, par exemple, $(x_1, \varphi(ih)) \in \Omega_h$, alors

$$-v_{\bar{x}_2 \hat{x}_2}(x_1, \varphi(ih)) = \left\{ \frac{v(x_1, \varphi(ih)) - v(x_1, \varphi((i-1)h))}{\varphi(ih) - \varphi((i-1)h)} - \right.$$

$$\left. - \frac{v(x_1, \varphi((i+1)h)) - v(x_1, \varphi(ih))}{\varphi((i+1)h) - \varphi(ih)} \right\} \frac{2}{\varphi((i+1)h) - \varphi((i-1)h)}.$$

On démontre l'unicité pour le problème algébrique non linéaire (4.14), (4.15) au moyen de l'estimation à priori du

THÉORÈME 4.2. *Soit u^h une solution du problème* (4.14), (4.15). *Alors toute fonction v définie sur $\bar{\Omega}_h$ et nulle sur Γ_h vérifie l'inégalité*

$$\max_{\Omega_h} |u^h(\mathbf{x}) - v(\mathbf{x})| \leq c_2 \delta,$$

où

$$\delta = \max_{\mathbf{x} \in \Omega_h} \left(|L^h v(\mathbf{x}) - f(v(\mathbf{x}), \mathbf{x})| (\min(x_2, b - x_2))^{2-\gamma} \right)$$

et la constante c_2 est indépendante de $\mathbf{x}$ *et h (mais elle dépend de $\gamma \in (0, 2\,])$.*

DÉMONSTRATION. Soit $w(\mathbf{x})$ solution du problème

$$L^h w(\mathbf{x}) = (\max\{x_2, b - x_2\})^{-2+\gamma}, \qquad \mathbf{x} \in \Omega_h,$$
$$w(\mathbf{x}) = 0, \qquad \mathbf{x} \in \Gamma_h. \tag{4.16}$$

On introduit une fonction discrète $\rho(\mathbf{x}) = \delta w(\mathbf{x})$. Le problème (4.16) satisfait au principe du maximum discret (voir [43]), et on a par un théorème de comparaison qui en découle:

$$w(\mathbf{x}) \geq 0 \qquad \forall\, \mathbf{x} \in \bar{\Omega}_h.$$

Comme $\delta \geq 0$, il en est de même de la fonction ρ:

$$\rho(\mathbf{x}) \geq 0 \qquad \forall\, \mathbf{x} \in \bar{\Omega}_h.$$

La fonction $f(p, \mathbf{x})$ admet des dérivées continues par rapport à p, d'où la validité du théorème de Lagrange

$$f(v(\mathbf{x}), \mathbf{x}) - f(u^h(\mathbf{x}), \mathbf{x}) = \frac{\partial f}{\partial p}(\xi(\mathbf{x}), \mathbf{x})\,(v(\mathbf{x}) - u^h(\mathbf{x}))$$

pour une fonction discrète $\xi(\mathbf{x})$ qui a toutes ses composantes entre $u^h(\mathbf{x})$ et $v(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \bar{\Omega}_h$.

Avec ces résultats, on transforme l'expression

$$\left(L^h - \frac{\partial f}{\partial p}(\xi(\mathbf{x}), \mathbf{x})\right)(v(\mathbf{x}) + \rho(\mathbf{x}) - u^h(\mathbf{x})) =$$
$$= L^h v(\mathbf{x}) - f(v(\mathbf{x}), \mathbf{x}) + \left(L^h - \frac{\partial f}{\partial p}(\xi(\mathbf{x}), \mathbf{x})\right)\rho(\mathbf{x}).$$

Du moment que $\partial f/\partial p$ est non positive, ρ est non négative, on a, compte tenu de (4.16),

$$\left(L^h - \frac{\partial f}{\partial p}(\xi(\mathbf{x}), \mathbf{x})\right)(v(\mathbf{x}) + \rho(\mathbf{x}) - u^h(\mathbf{x})) \geqslant$$
$$\geqslant L^h v(\mathbf{x}) - f(v(\mathbf{x}), \mathbf{x}) + \delta(\min\{x_2, b - x_2\})^{-2+\gamma} \geqslant 0.$$

La dernière inégalité découle de ce que

$$\delta \geqslant -(L^h v(\mathbf{x})) - f(v(\mathbf{x}), \mathbf{x}))(\min\{x_2, b - x_2\})^{2-\gamma}$$

par suite de la définition de δ.

On a sur Γ_h

$$v(\mathbf{x}) - u^h(\mathbf{x}) + \rho(\mathbf{x}) = 0.$$

La dérivée $\partial f/\partial p$ étant non positive, l'opérateur aux différences vérifie le principe du maximum. Aussi

$$v(\mathbf{x}) - u^h(\mathbf{x}) + \rho(\mathbf{x}) \geqslant 0 \quad \text{sur} \quad \bar{\Omega}_h.$$

On établit de même

$$u^h(\mathbf{x}) - v(\mathbf{x}) + \rho(\mathbf{x}) \geqslant 0 \quad \text{sur} \quad \bar{\Omega}_h.$$

Les deux inégalités impliquent évidemment

$$|u^h(\mathbf{x}) - v(\mathbf{x})| \leqslant \rho(\mathbf{x}) = \delta w(\mathbf{x}),$$

donc

$$\max_{\bar{\Omega}_h} |u^h(\mathbf{x}) - v(\mathbf{x})| \leqslant \delta \max_{\bar{\Omega}_h} |w(\mathbf{x})|. \tag{4.17}$$

On évalue $\max |w(\mathbf{x})|$ en appliquant une fois de plus le théorème de Lagrange à l'opérateur L^h. Soit $\gamma \in (0, 1)$. On introduit la fonction

$$z(x_1, x_2) = \frac{4^{2-\gamma}}{\gamma(1-\gamma)}(x_2^\gamma + (b - x_2)^\gamma)$$

et la fonction discrète

$$Z(\mathbf{x}) = z(\mathbf{x}), \qquad \mathbf{x} \in \bar{\Omega}_h.$$

On calcule $L^h Z$. On rappelle que la fonction g continue sur le segment $[t - h_1, t + h_2]$ et deux fois continûment dérivable sur

$(t-h_1, t+h_2)$ admet le développement taylorien et vérifie le théorème de Lagrange pour la dérivée d'ordre 2, d'où

$$\left(\frac{g(t+h_2)-g(t)}{h_2}-\frac{g(t)-g(t-h_1)}{h_1}\right)\frac{2}{h_1+h_2}=g''(\eta),$$

avec $\eta\in(t-h_1,\ t+h_2)$. Cette formule conduit à

$$Z_{\bar{x}_1 \hat{x}_1}(\mathbf{x})=0, \qquad \mathbf{x}\in\Omega_h,$$

$$Z_{\bar{x}_2\hat{x}_2}(x_1, \varphi(ih))=\frac{\partial^2 z}{\partial x_2^2}(x_1, \eta_i),$$

où $\eta_i\in\big(\varphi((i-1)h),\ \varphi((i+1)h)\big)$. Aussi

$$L^h Z(x_1, \varphi(ih))=-\frac{\partial^2 z}{\partial x_2^2}(x_1, \eta_1)=4^{2-\gamma}\left(\eta_i^{-2+\gamma}+(b-\eta_i)^{-2+\gamma}\right).$$

Les fonctions $x_2^{-2+\gamma}$ et $(b-x_2)^{-2+\gamma}$ étant monotones, on a

$$L^h Z(x_1, \varphi(ih))\geqslant 4^{2-\gamma}\left(\varphi^{-2+\gamma}((i+1)h)+(b-\varphi((i-1)h))^{-2+\gamma}\right).$$

On demande les valeurs extrémales de

$$\frac{\varphi(t+h)}{\varphi(t)}=\frac{3(t+h)^2-2(t+h)^3}{3t^2-2t^3},$$

avec $t\in[h, 1-h]$. Le maximum et le minimum réalisés aux extrémités du segment $[h, 1-h]$ sont respectivement

$$(12h^2-16h^3)/(3h^2-2h^3)\leqslant 4, \qquad (3h^2-2h^3)/(12h^2-16h^3)\geqslant 1/4.$$

Donc

$$L^h Z(x_1, x_2)\geqslant x_2^{-2+\gamma}+(b-x_2)^{-2+\gamma}\geqslant\left(\min\{x_2, b-x_2\}\right)^{-2+\gamma}.$$

En outre,

$$Z(\mathbf{x})\geqslant w(\mathbf{x})=0 \qquad \forall\,\mathbf{x}\in\Gamma_h.$$

On a par un théorème de comparaison

$$Z(\mathbf{x})=z(\mathbf{x})\geqslant w(\mathbf{x}) \qquad \forall\,\mathbf{x}\in\bar{\Omega}_h.$$

Comme

$$w(\mathbf{x})\leqslant\max_{\bar{\Omega}_h} z(\mathbf{x})\leqslant\frac{b^\gamma 2^{5-3\gamma}}{\gamma(1-\gamma)},$$

(4.17) donne le résultat du théorème à condition de poser

$$c_2=\frac{b^\gamma 2^{5-3\gamma}}{\gamma(1-\gamma)}.$$

Si $\gamma \in [1, 2]$, l'affirmation découle de ce qui précède car

$$(\min\{x_2, b - x_2\})^{\mu_1} \leqslant (b/2)^{\mu_1-\mu_2} (\min\{x_2, b - x_2\})^{\mu_2}, \quad \mu_1 \geqslant \mu_2 \geqslant 0.$$

Le théorème entraîne l'unicité de la solution du problème (4.14), (4.15). Soit, par exemple, u^h et v^h deux solutions distinctes de ce problème. Dans ce cas, on a l'inégalité du théorème 4.2, où $\delta = 0$. Du moment que la différence $u^h(\mathbf{x}) - v^h(\mathbf{x})$ est non nulle au moins en un point, on a

$$\max_{\Omega_h} |u^h(\mathbf{x}) - v^h(\mathbf{x})| > 0,$$

i.e. il y a contradiction.

On montre la possibilité du problème (4.14), (4.15) moyennant l'application de l'espace vectoriel des vecteurs ξ de composantes $\xi(\mathbf{x})$, $\mathbf{x} \in \Omega_h$, dans lui-même qui obéit à la formule

$$P: \xi(\mathbf{x}) \to \{L^h \xi(\mathbf{x}) - f(\xi(\mathbf{x}), \mathbf{x})\} \hbar_1(\mathbf{x}) \hbar_2(\mathbf{x}), \qquad \mathbf{x} \in \Omega_h.$$

(Pour des raisons d'homogénéité, on a posé $\xi(\mathbf{x}) = 0$ sur Γ_h.) Quel est le sens des symboles $\hbar_i(\mathbf{x})$? Soit $\mathbf{x} = (x_1, x_2) \in \Omega_h$, auquel cas ce point est avoisiné sur $\bar{\Omega}_h$ à gauche par $(x_1 - \mu_1, x_2)$, à droite par $(x_1 + \mu_2, x_2)$, en haut par $(x_1, x_2 + \mu_3)$ et en bas par $(x_1, x_2 - \mu_4)$. On pose

$$\hbar_1(\mathbf{x}) = \frac{\mu_1 + \mu_2}{2}, \qquad \hbar_2(\mathbf{x}) = \frac{\mu_3 + \mu_4}{2},$$

$$\hbar_1^+(\mathbf{x}) = \mu_2, \qquad \hbar_1^-(\mathbf{x}) = \mu_1, \qquad \hbar_2^+(\mathbf{x}) = \mu_3, \qquad \hbar_2^-(\mathbf{x}) = \mu_4.$$

On note la continuité de l'application P. On utilise le lemme 6.1, § 3.6, et on calcule à cet effet la somme

$$\sum_{\mathbf{x} \in \Omega_h} \xi(\mathbf{x}) P\xi(\mathbf{x}) = $$
$$= \sum_{\mathbf{x} \in \Omega_h} \{-\xi_{\bar{x}_1 \hat{x}_1}(\mathbf{x}) - \xi_{\bar{x}_2 \hat{x}_2}(\mathbf{x}) - f(\xi(\mathbf{x}), \mathbf{x})\} \xi(\mathbf{x}) \hbar_1(\mathbf{x}) \hbar_2(\mathbf{x}). \tag{4.18}$$

Chaque terme est étudié séparément. On applique pour x_1 fixé la première formule de Green discrète (voir [42]), il vient

$$-\sum_{\substack{\mathbf{x} \in \Omega_h \\ x_1 = \text{const}}} \xi_{\bar{x}_2 \hat{x}_2}(\mathbf{x}) \xi(\mathbf{x}) \hbar_1(\mathbf{x}) \hbar_2(\mathbf{x}) =$$
$$= -\sum_{i=1}^{k} \xi_{\bar{x}_2 \hat{x}_2}(x_1, \varphi(ih)) \xi(x_1, \varphi(ih)) \hbar_1(\mathbf{x}) \hbar_2(\mathbf{x}) =$$
$$= \sum_{i=0}^{k} \left(\frac{\xi(x_1, \varphi((i+1)h)) - \xi(x_1, \varphi(ih))}{\hbar_2^+(\mathbf{x})} \right)^2 \hbar_1(\mathbf{x}) \hbar_2^+(\mathbf{x}).$$

Dans cette égalité, k est choisi à partir de la condition $(x_1, \varphi((k+1)h)) \in \Gamma_h$, si bien que

$$\xi(x_1, 0) = 0, \qquad \xi(x_1, \varphi((k+1)h)) = 0.$$

On établit la minoration à l'aide de l'inégalité de Friedrichs discrète (voir [43]) :

$$-\sum_{\substack{\mathbf{x}\in\Omega_h \\ x_1=\mathrm{const}}} \xi_{\bar{x}_2 x_2}(\mathbf{x})\,\xi(\mathbf{x})\,\hbar_1(\mathbf{x})\,\hbar_2(\mathbf{x}) \geqslant$$

$$\geqslant \frac{4}{\varphi^2((k+1)h)} \sum_{i=1}^{k} \xi^2(x_1, \varphi(ih))\,\hbar_1(\mathbf{x})\,h_2(\mathbf{x}) \geqslant$$

$$\geqslant \frac{4}{b^2} \sum_{\substack{\mathbf{x}\in\Omega_h \\ x_1=\mathrm{const}}}^{k} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x})$$

On somme par rapport aux valeurs de x_1 qui constituent le réseau de discrétisation, ce qui donne

$$-\sum_{\mathbf{x}\in\Omega_h} \xi_{\bar{x}_2 x_2}(\mathbf{x})\,\xi(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}) \geqslant \frac{4}{b^2} \sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}). \qquad (4.19)$$

On démontre de même l'inégalité

$$-\sum_{\mathbf{x}\in\Omega_h} \xi_{\bar{x}_1 x_1}(\mathbf{x})\,\xi(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}) \geqslant \frac{4}{a_1^2} \sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}). \qquad (4.20)$$

Le troisième terme de (4.18) est transformé comme suit :

$$-\sum_{\mathbf{x}\in\Omega_h} f(\xi(\mathbf{x}), \mathbf{x})\,\xi(\mathbf{x})\,h_1(\mathbf{x})\,\hbar_2(\mathbf{x}) =$$

$$= -\sum_{\mathbf{x}\in\Omega_h} f(0, x)\,\xi(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}) -$$

$$-\sum_{\mathbf{x}\in\Omega_h} \{f(\xi(\mathbf{x}), \mathbf{x}) - f(0, \mathbf{x})\}\,\xi(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}) \geqslant$$

$$\geqslant -\max_{\bar{\Omega}} |f(0, \mathbf{x})| \sum_{\mathbf{x}\in\Omega_h} |\xi(\mathbf{x})|\,h_1(\mathbf{x})\,h_2(\mathbf{x}) -$$

$$-\sum_{\mathbf{x}\in\Omega_h} \frac{\partial f}{\partial p}\,\eta(\mathbf{x}), \mathbf{x})\,\xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}).$$

Ici $\eta(\mathbf{x})$ est situé entre 0 et $\xi(\mathbf{x})\ \forall\, \mathbf{x}\in\Omega_h$. On applique au premier terme du second membre l'inégalité de Cauchy-Bouniakovski tout en rejetant le deuxième:

$$\sum_{\mathbf{x}\in\Omega_h} f(\xi(\mathbf{x}),\ \mathbf{x})\,\xi(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}) \geq$$

$$\geq -\max_{\mathbf{x}\in\bar{\Omega}} |f(0,\ \mathbf{x})| \left(\sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x})\right)^{1/2} \times$$

$$\times \left(\sum_{\mathbf{x}\in\Omega_h} h_1(\mathbf{x})\,h_2(\mathbf{x})\right)^{1/2} \geq -c_3\left(\sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x})\right)^{1/2}, \tag{4.21}$$

où

$$c_3 = \max_{\mathbf{x}\in\bar{\Omega}} |f(0,\ \mathbf{x})| \left(\frac{a_2 b}{2}\right)^{1/2}.$$

Les inégalités (4.19) à (4.21) donnent

$$\sum_{\mathbf{x}\in\Omega_h} \xi(\mathbf{x})\,P_\xi(\mathbf{x}) \geq \left(\frac{4}{b^2}+\frac{4}{a_1^2}\right) \sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x}) -$$

$$- c_3\left(\sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x})\right)^{1/2}.$$

Quel que soit

$$\sigma \geq \sigma_0 = \frac{c_3}{\dfrac{4}{b^2}+\dfrac{4}{a_1^2}},$$

on a

$$\left(\frac{4}{b^2}+\frac{4}{a_1^2}\right)\sigma^2 - c_3\sigma \geq 0.$$

Aussi on a, pour toute fonction discrète $\xi(\mathbf{x})$ définie sur Ω_h telle que

$$\left(\sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\right)^{1/2} = \frac{\sigma_0}{\left(\min\limits_{\mathbf{x}\in\Omega_h} h_1(\mathbf{x})\,h_2(\mathbf{x})\right)^{1/2}},$$

l'estimation

$$\left(\sum_{\mathbf{x}\in\Omega_h} \xi^2(\mathbf{x})\,h_1(\mathbf{x})\,h_2(\mathbf{x})\right)^{1/2} \geq \sigma_0,$$

donc

$$\sum_{\mathbf{x}\in\Omega_h} \xi(\mathbf{x})\,P_\xi(\mathbf{x}) \geq 0.$$

Avec ces inégalités, le problème

$$P_\xi(\mathbf{x}) = 0, \qquad \mathbf{x} \in \Omega_h,$$

possède par le lemme 6.1, § 3.6 une solution $\xi(\mathbf{x})$ vérifiant la condition aux limites homogène

$$\xi(\mathbf{x}) = 0 \quad \text{sur} \quad \Gamma_h.$$

D'où la possibilité du problème (4.14), (4.15). Il y a plus. Le théorème 4.2 implique non seulement l'existence et l'unicité, mais aussi une estimation du module de la solution. Si l'on pose $v(\mathbf{x}) \equiv 0$, on a pour tout $\gamma \in (0, 2]$ fixé

$$\max_{\bar{\Omega}_h} |u^h(\mathbf{x})| \leqslant c_2 \max_{\bar{\Omega}} |d^{2-\gamma}(\mathbf{x}) f(0, \mathbf{x})|. \tag{4.22}$$

On suppose maintenant que la fonction $v(\mathbf{x})$ du théorème 4.2 est remplacée par la solution du problème (4.1) à (4.4). On calcule la différence

$$L^h u(\mathbf{x}) - f(u(\mathbf{x}), \mathbf{x}).$$

Soit $\mathbf{x} = (x_1, \varphi(ih)) \in \Omega_h$, $i \neq 1$, auquel cas

$$u(x_1, \varphi((i+1)h)) = u(\mathbf{x}) + h_2^+(\mathbf{x}) \frac{\partial u}{\partial x_2}(\mathbf{x}) +$$
$$+ \frac{1}{2}(h_2^+(\mathbf{x}))^2 \frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) + \frac{1}{6}(h_2^+(\mathbf{x}))^3 \frac{\partial^3 u}{\partial x_2^3}(\mathbf{x}) + \frac{1}{6}(h_2^+(\mathbf{x}))^{3+\alpha}\xi^+(\mathbf{x}).$$

Les majorations (4.13) entraînent

$$|\xi^+(\mathbf{x})| \leqslant \frac{\left|\frac{\partial^3 u}{\partial x_2^3}(\mathbf{x}) - \frac{\partial^3 u}{\partial x_2^3}(\mathbf{x}^+)\right|}{|\mathbf{x} - \mathbf{x}^+|^\alpha} \leqslant c_1 (\min\{d(\mathbf{x}), d(\mathbf{x}^+)\})^{-2-\alpha}.$$

Le point $\mathbf{x}^+$ est sur le segment de droite joignant les nœuds $\mathbf{x} = (x_1, \varphi(ih))$ et $(x_1, \varphi((i+1)h))$. Donc

$$|\xi^+(\mathbf{x})| \leqslant c_1 (\min\{\varphi((i+1)h), \varphi((i-1)h),$$
$$b - \varphi((i+1)h), b - \varphi((i-1)h)\})^{-2-\alpha}.$$

On a établi au cours de la démonstration du théorème 4.2 que

$$\varphi((i \pm 1)h) \leqslant 4\varphi(ih), \qquad |b - \varphi((i \pm 1)h)| \leqslant$$
$$\leqslant 4(b - \varphi(ih)). \tag{4.23}$$

C'est pourquoi

$$|\xi^+(\mathbf{x})| \leqslant 4^{2+\alpha} c_1 (\min(x_2, b - x_2))^{-2-\alpha}. \tag{4.24}$$

La relation

$$u(x_1, \varphi((i-1)h)) = u(\mathbf{x}) - h_2^-(\mathbf{x})\frac{\partial u}{\partial x_2}(\mathbf{x}) + \\ + \frac{1}{2}(h_2^-(\mathbf{x}))^2\frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) - \frac{1}{6}(h_2^-(\mathbf{x}))^3\frac{\partial^3 u}{\partial x_2^3}(\mathbf{x}) + \\ + \frac{1}{6}(h_2^-(\mathbf{x}))^{3+\alpha}\xi^-(\mathbf{x}) \quad (4.25)$$

et l'estimation

$$|\xi^-(\mathbf{x})| \leqslant 4^{2+\alpha}c_1(\min\{x_2, b-x_2\})^{-2-\alpha}$$

sont obtenues de façon analogue. Les deux développements donnent la formule

$$-u_{\bar{x}_2 \hat{x}_2}(\mathbf{x}) = -\frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) - \frac{h_2^+(\mathbf{x}) - h_2^-(\mathbf{x})}{3}\frac{\partial^3 u}{\partial x_2^3}(\mathbf{x}) - \\ - \frac{(h_2^+(\mathbf{x}))^{2+\alpha}\xi^+(\mathbf{x}) + (h_2^-(\mathbf{x}))^{2+\alpha}\xi^-(\mathbf{x})}{3(h_2^+(\mathbf{x}) + h_2^-(\mathbf{x}))}. \quad (4.26)$$

On a, en vertu de la définition de la fonction φ,

$$h_2^+(\mathbf{x}) - h_2^-(\mathbf{x}) = \varphi((i+1)h) - 2\varphi(ih) + \varphi((i-1)h) = \\ = bh^2(6 - 12\eta_i),$$

où $\eta_i \in ((i-1)h, (i+1)h)$; donc

$$\frac{|h_2^+(\mathbf{x}) - h_2^-(\mathbf{x})|}{3} \leqslant 2bh^2,$$

mais $\varphi(h) = b - \varphi(1-h) \geqslant 2bh^2$, si bien que

$$\frac{|h_2^+(\mathbf{x}) - h_2^-(\mathbf{x})|}{3} \leqslant h^{1+\alpha}(\min\{x_2, b-x_2\})^{\frac{1-\alpha}{2}}(2b)^{\frac{1+\alpha}{2}}. \quad (4.27)$$

Les majorations (4.13) impliquent également l'inégalité

$$\left|\frac{\partial^3 u}{\partial x_2^3}(\mathbf{x})\right| \leqslant c_1 d^{-2}(\mathbf{x}). \quad (4.28)$$

On a enfin dans le second membre de l'égalité (4.26)

$$\frac{(h_2^+(\mathbf{x}))^{2+\alpha}}{h_2^+(\mathbf{x}) + h_2^-(\mathbf{x})} \leqslant (h_2^+(\mathbf{x}))^{1+\alpha} = h^{1+\alpha}|\varphi'(\xi_i)|^{1+\alpha},$$

avec $\xi_i \in (ih, (i+1)h)$. Soit $t \leqslant 1/2$. Comme

$$\varphi'(t) = 6b(t - t^2),$$

on a

$$|\varphi'(t)| = 6bt(1-t) \leqslant 6bt.$$

D'autre part,

$$\varphi(t) = b(3t^2 - 2t^3) \geqslant 2t^2 b.$$

Aussi

$$|\varphi'(t)| \leqslant 3\sqrt{2b}(\varphi(t))^{1/2}.$$

Si $t \geqslant 1/2$, on trouve de même

$$|\varphi'(t)| \leqslant 3\sqrt{2b}(b - \varphi(t))^{1/2}.$$

Si l'on prend en considération (4.23), alors

$$\varphi'(\xi_i)| \leqslant 3\sqrt{2b}(\min\{\varphi(\xi_i), b - \varphi(\xi_i)\})^{1/2} \leqslant$$
$$\leqslant 6\sqrt{2b}(\min\{x_2, b - x_2\})^{1/2}. \quad (4.29)$$

S'agissant de l'égalité (4.26), les relations (4.24), (4.25), (4.27) à (4.29) donnent

$$\left| -u_{\bar{x}_2 \hat{x}_2}(\mathbf{x}) + \frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) \right| \leqslant$$
$$\leqslant h^{1+\alpha} c_1 d^{-2}(\mathbf{x}) (\min\{x_2, b - x_2\})^{\frac{1-\alpha}{2}} (2b)^{\frac{\alpha+1}{2}} +$$
$$+ h^{1+\alpha}\left(6\sqrt{2b}(\min\{x_2, b - x_2\})^{1/2}\right)^{1+\alpha} \cdot 4^{2+\alpha} c_1 (\min\{x_2, b -$$
$$- x_2\})^{-2-\alpha} \leqslant c_4 h^{1+\alpha}(\min\{x_2, b - x_2\})^{\frac{-3-\alpha}{2}}. \quad (4.30)$$

On a établi ces inégalités sous l'hypothèse de $\mathbf{x} = (x_1, \varphi(ih)) \in \Omega_h$, $i \neq 1$. Soit $i = 1$. Comme $u \in C^1(\bar{\Omega})$, on a

$$u(x_1, 0) = u(\mathbf{x}) - \varphi(h)\zeta^-(\mathbf{x})$$

et

$$u(x_1, \varphi(2h)) = u(\mathbf{x}) + (\varphi(2h) - \varphi(h))\zeta^+(\mathbf{x}),$$

où

$$|\zeta^{\pm}(\mathbf{x})| \leqslant c_5.$$

On porte ces développements dans la différence divisée seconde, il vient

$$\left| u_{\bar{x}_2\bar{x}_2}(\mathbf{x}) \right| = \frac{2}{\varphi(2h)} \left| \zeta^+(\mathbf{x}) - \zeta^-(\mathbf{x}) \right| \leqslant \frac{4c_3}{\varphi(2h)}.$$

La fonction φ vérifie les inégalités $\varphi(2h) \geqslant \varphi(h)$ et $\varphi(h) \leqslant 3bh^2$, qui impliquent

$$\left| u_{\bar{x}_2\bar{x}_2}(\mathbf{x}) \right| \leqslant \frac{4c_3}{\varphi(h)} \leqslant c_6 h^{1+\alpha} \varphi^{(-3-\alpha)/2}(h).$$

On rappelle que $x_2 = \varphi(h)$; aussi il résulte de (4.13) la majoration

$$\left| \frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) \right| \leqslant c_1 d^{-1}(\mathbf{x}) \leqslant \frac{c_1}{\varphi(h)} \leqslant c_7 h^{1+\alpha} \varphi^{(-3-\alpha)/2}(h).$$

Les deux estimations obtenues donnent

$$\left| - u_{\bar{x}_2\bar{x}_2}(\mathbf{x}) + \frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) \right| \leqslant c_8 h^{1+\alpha} \varphi^{(-3-\alpha)/2}(h) \qquad (4.31)$$

à condition que $x_2 = \varphi(h)$.

Soit $c_9 = \max(c_4, c_8)$. Les inégalités (4.30), (4.31) permettent d'écrire

$$\left| - u_{\bar{x}_2\bar{x}_2}(\mathbf{x}) + \frac{\partial^2 u}{\partial x_2^2}(\mathbf{x}) \right| \leqslant$$

$$\leqslant c_9 h^{1+\alpha} (\min\{x_2, b - x_2\})^{(-3-c)/2} \qquad \forall\, \mathbf{x} \in \Omega_h. \quad (4.32)$$

S'agissant de l'autre variable, on trouve par des raisonnements analogues

$$\left| - u_{\bar{x}_1\bar{x}_1}(\mathbf{x}) + \frac{\partial^2 u}{\partial x_1^2}(\mathbf{x}) \right| \leqslant$$

$$\leqslant c_{10} h^{1+\alpha} (\min\{x_2, b - x_2\})^{(-3-\alpha)/2} \qquad \forall\, \mathbf{x} \in \Omega_h. \quad (4.33)$$

On a finalement en réunissant (4.32) et (4.33) :

$$\left| L^h u(\mathbf{x}) + \Delta u(\mathbf{x}) \right| \leqslant c_{11} h^{1+\alpha} (\min\{x_2, b - x_2\})^{(-3-\alpha)/2} \qquad \forall\, \mathbf{x} \in \Omega_h.$$

L'équation (4.1) entraîne $\Delta u = -f(u(\mathbf{x}), \mathbf{x})$, si bien que

$$\left| L^h u(\mathbf{x}) - f(u(\mathbf{x}), \mathbf{x}) \right| \leqslant c_{11} h^{1+\alpha} (\min\{x_2, b - x_2\})^{(-3-\alpha)/2} \qquad \forall\, \mathbf{x} \in \Omega_h.$$

On en tire, en vertu de l'estimation du théorème 4.2, le résultat suivant.

THÉORÈME 4.3. *On suppose que le problème* (4.1), (4.3) *vérifie les conditions* (4.2), (4.5). *La solution* u^h *du problème* (4.14), (4.15) *admet l'estimation*

$$\max_{\mathbf{x}\in\bar{\Omega}_h} \left| u^h(\mathbf{x}) - u(\mathbf{x}) \right| \leqslant c_{12} h^{1+\alpha},$$

avec α *un nombre fixé de l'intervalle* (0, 1).

On note l'impossibilité de passer à la limite pour $\alpha \to 1$. En effet, les raisonnements ci-dessus impliquent la croissance de c_{12} avec α tendant vers 1.

4.5. Sur le problème de diffraction

Le problème de diffraction conduit au problème aux limites pour une équation dont les coefficients subissent des discontinuités de première espèce. On pose aux lignes de discontinuité les conditions de transmission qui sont déterminées par la continuité de la matière et l'équilibre des forces agissantes. La discontinuité des coefficients correspond à la présence dans le milieu de deux ou plusieurs corps de caractéristiques physiques différentes.

Si les lignes de discontinuité sont régulières sans se recouper et rencontrer la frontière, la solution n'est plus régulière globalement, i.e. dans l'ensemble du domaine. Mais elle le reste dans chaque domaine partiel limité par les lignes de discontinuité, cette propriété ayant lieu jusques et y compris ces lignes. Si celles-ci aboutissent à la frontière, la régularité est altérée comme dans le cas de points anguleux, ce qui exige des fois des schémas variationnels aux différences de forme particulière. Nous étudierons pour notre part un cas où l'on atteint une précision suffisante à l'aide d'un schéma usuel, et nous utiliserons, pour élever l'ordre de précision, un procédé simple relevant du passage d'une norme de l'erreur à une norme plus faible.

Soit le problème de diffraction simple dans lequel le milieu est formé de deux corps différents. On cherche la solution de l'équation

$$-\frac{\partial}{\partial x_1} a \frac{\partial u}{\partial x_1} - \frac{\partial}{\partial x_2} a \frac{\partial u}{\partial x_2} = f \tag{5.1}$$

dans le rectangle $\Omega = \{(x_1, x_2): 0 < x_1 < b_1;\ 0 < x_2 < b_2\}$ avec la condition à la frontière Γ:

$$u(\mathbf{x}) = 0, \qquad \mathbf{x} \in \Gamma. \tag{5.2}$$

La droite $x_1 = b_1/2$ partage Ω en deux rectangles

$$\Omega_1 = \{(x_1, x_2);\ 0 < x_1 < b_1/2,\ 0 < x_2 < b_2\},$$

$$\Omega_2 = \{(x_1, x_2);\ b_1/2 < x_1 < b_1,\ 0 < x_2 < b_2\},$$

de frontière commune S. On suppose que le coefficient $a(\mathbf{x})$ est constant par morceaux:

$$a(\mathbf{x}) = a_i > 0 \quad \text{pour} \quad \mathbf{x} \in \Omega_i,$$

$$f(\mathbf{x}) = f_i(\mathbf{x}).$$

On a en fait deux équations dans deux domaines. Pour qu'il y ait existence et unicité, il faut poser des conditions supplémentaires à l'interface S. On admet que la solution vérifie deux conditions de transmission

$$[u(\mathbf{x})]_S = 0, \quad \mathbf{x} \in S, \tag{5.3}$$

$$\left[a(\mathbf{x}) \frac{\partial u}{\partial x_1}(\mathbf{x})\right]_S = 0, \quad \mathbf{x} \in S. \tag{5.4}$$

Le symbole $[\varphi(\mathbf{x})]_S$ signifie la différence des valeurs limites que φ prend en $\mathbf{x}$ de part et d'autre de S. On suppose que

$$f \in C^\alpha(\bar{\Omega}_1) \cap C^\alpha(\bar{\Omega}_2), \qquad \alpha \in (0, 1). \tag{5.5}$$

Avec ces conditions (voir [103]), la solution u existe, est unique, appartient dans Ω_1 et Ω_2 à $C^{2+\alpha}(\Omega_i)$, et elle n'est en général pas dans $C^{2+\alpha}(\bar{\Omega}_i)$. Le problème ainsi posé équivaut à chercher une fonction $u \in \mathring{W}_2^1(\Omega)$ satisfaisant à l'identité intégrale

$$\int_\Omega a \sum_{i=1}^{2} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, d\mathbf{x} = \int_\Omega f v \, d\mathbf{x} \qquad \forall v \in \mathring{W}_2^1(\Omega). \tag{5.6}$$

Aux termes du théorème 1.3, les conditions énumérées garantissent, pour le problème (5.6), l'existence d'une solution unique dans la classe $\mathring{W}_2^1(\Omega)$. On établit de plus dans [103] une régularité plus grande de la solution u. Premièrement, les propriétés générales d'une solution généralisée impliquent que $u \in W_2^2(\Omega_i')$ pour tout domaine $\Omega_i' \subset \Omega_i$, $i = 1, 2$, tel que la distance de Ω_i' à S soit positive. Cela étant, on a l'estimation

$$\|u\|_{W_2^2(\Omega_i')} \leq c_1 \|f\|_{L_2(\Omega_i)}, \tag{5.7}$$

avec la constante c_1 indépendante de f et u, mais dépendant de Ω_i'. Deuxièmement, le fait d'utiliser les conditions de transmission

(5.3). (5.4) permet de dire que $u \in W_2^2(\Omega_i \cap \Omega')$, $i = 1, 2$, pour tout $\Omega' \subset \Omega$ dont la distance à Γ est un nombre positif, et

$$\|u\|_{W_2^2(\Omega' \cap \Omega_1)} + \|u\|_{W_2^2(\Omega' \cap \Omega_2)} \leqslant c_2 \|f\|_{L_2(\Omega)}. \qquad (5.8)$$

Il nous reste à étudier u de classe W_2^2 au voisinage des points d'intersection de S et Γ.

A cet effet, on utilise les procédés décrits dans le paragraphe précédent pour les points anguleux (on introduit au voisinage des points singuliers les coordonnées polaires et on cherche la solution comme série de Fourier). On note que la solution et le second membre se développent dans une région autour du point singulier en séries procédant suivant un système spécial de fonctions régulières par morceaux (voir [38]) pour le produit scalaire avec poids constant par intervalles. Si l'on examine en détail la convergence des séries, on constate une particularité de l'angle $\pi/2$ sous lequel la frontière Γ coupe S, à savoir la solution u du problème de diffraction appartient dans chaque Ω_i à $W_2^2(\Omega_i)$. Cette régularité suffit pour que la méthode des éléments finis procure une précision suffisante.

La motivation étant trop laborieuse, on démontrera cette affirmation de façon plus simple. Soit le rectangle $\Omega^S = \{(x_1, x_2);\ 0 < x_1 < b_1, -b_2 < x_2 < b_2\}$ contenant le domaine Ω. On prolonge a et f de Ω à Ω^S en fonctions paire et impaire respectivement par rapport à l'axe Ox_1:

$$\tilde{a}(\mathbf{x}) = \begin{cases} a_1 & \text{si} \quad x_1 < b_1/2, \\ a_2 & \text{si} \quad x_1 > b_1/2, \end{cases}$$

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(x_1, x_2) & \text{si} \quad x_2 \geqslant 0, \\ -f(x_1, -x_2) & \text{si} \quad x_2 < 0. \end{cases} \qquad (5.9)$$

On note que $\tilde{f} \in L_2(\Omega^S)$ et $\|\tilde{f}\|_{L_2(\Omega^S)} \leqslant \sqrt{2}\, \|f\|_{L_2(\Omega)}$.

Soit le problème de trouver une fonction $\tilde{u} \in \overset{\circ}{W}{}_2^1(\Omega^S)$ vérifiant l'identité intégrale

$$\int_{\Omega^S} \tilde{a} \sum_{i=1}^{2} \frac{\partial \tilde{u}}{\partial x_i} \frac{\partial \tilde{v}}{\partial x_i}\, d\mathbf{x} = \int_{\Omega^S} \tilde{f}\, \tilde{v}\, d\mathbf{x} \qquad \forall\, \tilde{v} \in \overset{\circ}{W}{}_2^1(\Omega^S). \qquad (5.10)$$

Selon le théorème 1.3, le problème admet une solution unique dans cette classe. La propriété mentionnée fait de plus qu'elle appartient à $W_2^2(\Omega_i \cap \Omega')$, $i = 1, 2$, pour tout $\Omega' \subset \Omega$ tel que la distance de Ω' et de la frontière de Ω^S (et non de Ω!) soit une quantité positive. Le domaine partiel Ω' remplit cette condition

s'il s'agit d'un voisinage suffisamment restreint du point $(b_1/2, 0)$. Cela étant, on a l'estimation

$$\|\tilde{u}\|_{W_2^2(\Omega'\cap\Omega_1)} + \|\tilde{u}\|_{W_2^2(\Omega'\cap\Omega_2)} \leq c_3\|\tilde{f}\|_{L_2(\Omega^S)} \leq \sqrt{2}\, c_3\|f\|_{L_2(\Omega)}. \quad (5.11)$$

On montre pour le problème (5.10) l'existence d'une solution

$$\tilde{\tilde{u}}(\mathbf{x}) = \begin{cases} u(x_1, x_2) & \text{si } x_2 \geq 0, \\ -u(x_1, -x_2) & \text{si } x_2 < 0, \end{cases} \quad (5.12)$$

qui est impaire par rapport à l'axe Ox_1. En effet, $u \in \overset{\circ}{W}{}_2^1(\Omega)$ entraîne $\tilde{\tilde{u}} \in \overset{\circ}{W}{}_2^1(\Omega^S)$. On l'établit le plus aisément à l'aide de la définition de l'espace $\overset{\circ}{W}{}_2^1$ donnée par [34]. Soit $\tilde{v} \in \overset{\circ}{W}{}_2^1(\Omega^S)$ quelconque. Cette fonction vérifie les égalités

$$\int_{\Omega^S} a \sum_{i=1}^{2} \frac{\partial\tilde{\tilde{u}}}{\partial x_i}\frac{\partial \tilde{v}}{\partial x_i}\,d\mathbf{x} = \int_{\substack{\Omega^S\\ x_2>0}} a \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}\frac{\partial \tilde{v}}{\partial x_i}\,d\mathbf{x} +$$

$$+ \int_{\substack{\Omega^S\\ x_2<0}} \tilde{a}\left(-\frac{\partial u}{\partial x_1}(x_1, -x_2)\frac{\partial \tilde{v}}{\partial x_1} + \frac{\partial u}{\partial x_2}(x_1, -x_2)\frac{\partial \tilde{v}}{\partial x_2}\right)d\mathbf{x} =$$

$$= \int_{\Omega} a \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}\frac{\partial \tilde{v}}{\partial x_i}\,d\mathbf{x} - \int_{\Omega} a\left(-\frac{\partial u}{\partial x_1}\frac{\partial \tilde{v}}{\partial x_1}(x_1, -x_2) +\right.$$

$$\left. + \frac{\partial u}{\partial x_2}\frac{\partial \tilde{v}}{\partial x_2}(x_1, -x_2)\right)d\mathbf{x} = \int_{\Omega} a \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}\frac{\partial w}{\partial x_i}\,d\mathbf{x},$$

avec w définie par

$$w(x_1, x_2) = \tilde{v}(x_1, x_2) - \tilde{v}(x_1, -x_2). \quad (5.13)$$

Avec la définition de [34], la propriété $v \in \overset{\circ}{W}{}_2^1(\Omega^S)$ implique $w \in \overset{\circ}{W}{}_2^1(\Omega)$. Aussi (5.6) donne lieu à l'égalité

$$\int_{\Omega^S} \tilde{a} \sum_{i=1}^{2} \frac{\partial\tilde{\tilde{u}}}{\partial x_i}\frac{\partial \tilde{v}}{\partial x_i}\,d\mathbf{x} = \int_{\Omega} f\tilde{v}\,d\mathbf{x} - \int_{\Omega} f\tilde{v}(x_1, -x_2)\,d\mathbf{x}.$$

En se rappelant la façon dont on a construit la fonction $\tilde{f}$, on a

$$\int_{\Omega^S} \tilde{a} \sum_{i=1}^{2} \frac{\partial\tilde{\tilde{u}}}{\partial x_i}\frac{\partial \tilde{v}}{\partial x_i}\,d\mathbf{x} = \int_{\Omega^S} \tilde{f}\tilde{v}\,d\mathbf{x}. \quad (5.14)$$

Comme la fonction $\tilde{v}$ a été prise quelconque, $\tilde{\tilde{u}}$ est solution du problème (5.10). Etant donnée l'unicité, la solution de (5.10) est définie par la formule (5.12) et jouit de la propriété (5.11).

Si l'on utilise les estimations (5.7), (5.8) et (5.11) à la fois, on établit que la solution u du problème (5.6) est dans $W_2^2(\Omega' \cap \Omega_i)$, $i = 1, 2$, pour tout $\Omega' \subset \Omega$ dont la distance au point unique $(b_1/2, b_2)$ est positive. Puisque les raisonnements précédents restent vrais pour la région autour de ce point, on a

$$u \in W_2^2(\Omega_1) \cap W_2^2(\Omega_2),$$

et

$$\|u\|_{W_2^2(\Omega_1)} + \|u\|_{W_2^2(\Omega_2)} \leqslant c_4 \|f\|_{L_2(\Omega)}. \tag{5.15}$$

Ainsi, on a établi des résultats de régularité pour la solution u. On se propose de construire un schéma variationnel aux différences par la méthode de Galerkin.

Soit le réseau rectangulaire uniforme

$$\bar{\Omega}_h = \{(x_1, x_2):\ x_1 = ih_1,\ x_2 = jh_2,\ i = 0, 1, \ldots, 2N_1,\ j = \\ = 0, 1, \ldots, N_2\}$$

de pas $h_1 = b_1/(2N_1)$, $h_2 = b_2/N_2$, où $N_i \geqslant 2$ sont des entiers. Pour que chaque maille du réseau soit entièrement dans un des domaines Ω_i, les intervalles partiels du segment de l'axe Ox_1 sont pris en nombre pair. On introduit la notation $\Omega_h = \bar{\Omega}_h \cap \Omega$ et on fait correspondre à chaque point $\mathbf{y} = (y_1, y_2) \in \Omega_h$ une fonction de base $\varphi_\mathbf{y}(\mathbf{x}) \in \overset{\circ}{W}{}_2^1(\Omega)$ qui est égale à 1 en $\mathbf{y}$, nulle partout ailleurs sur $\bar{\Omega}_h$ et linéaire par morceaux sur chaque triangle élémentaire ouvert (fig. 4.7) déterminé par les diagonales des mailles rectangulaires, qui forment un angle aigu avec l'axe Oy_1 (fig. 4.8). Ainsi, on trouve:

$$\varphi_\mathbf{y}(\mathbf{x}) = \begin{cases} 1 + \dfrac{1}{h_2}(y_2 - x_2), & \mathbf{x} \in \overline{T}_1, \\ 1 + \dfrac{1}{h_1}(y_1 - x_1), & \mathbf{x} \in \overline{T}_2, \\ 1 + \dfrac{1}{h_1}(y_1 - x_1) - \dfrac{1}{h_2}(y_2 - x_2), & \mathbf{x} \in \overline{T}_3, \\ 1 - \dfrac{1}{h_2}(y_2 - x_2), & \mathbf{x} \in \overline{T}_4, \\ 1 - \dfrac{1}{h_1}(y_1 - x_1), & \mathbf{x} \in \overline{T}_5, \\ 1 - \dfrac{1}{h_1}(y_1 - x_1) + \dfrac{1}{h_2}(y_2 - x_2), & \mathbf{x} \in \overline{T}_6, \\ 0 \text{ dans les cas restants.} \end{cases} \tag{5.16}$$

On désigne par H^h le sous-espace vectoriel engendré par ce système de fonctions. Il est clair que $H^h \subset \overset{\circ}{W}{}_2^1(\Omega)$.

Conformément au principe de la méthode des éléments finis (voir [112], [132]), on cherche la solution approchée $u^h \in H^h$ sous forme de somme

$$u^h(\mathbf{x}) = \sum_{\mathbf{y} \in \Omega_h} \alpha_{\mathbf{y}} \varphi_{\mathbf{y}}(\mathbf{x}), \tag{5.17}$$

où $\{\alpha_{\mathbf{y}}\}$ est un jeu de constantes définies par les égalités obtenues à partir de (5.6) par les substitutions $u = u^h$ et $v = \varphi_{\mathbf{y}}$:

$$L(u^h, \varphi_{\mathbf{z}}) = (f, \varphi_{\mathbf{z}}) \qquad \forall \mathbf{z} \in \Omega_h. \tag{5.18}$$

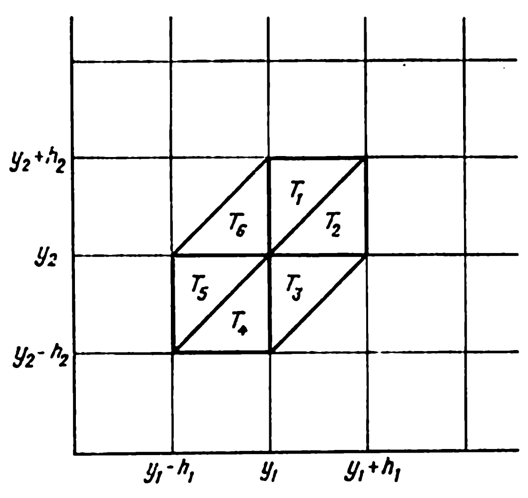


Fig. 4.7. Triangulation du domaine

Fig. 4.8. Fonction de base

Ici le produit scalaire et la forme bilinéaire sont définis par les formules

$$(v, w) = \int_{\Omega} vw \, d\mathbf{x},$$

$$L(v, w) = \int_{\Omega} a \left(\frac{\partial v}{\partial x_1} \frac{\partial w}{\partial x_1} + \frac{\partial v}{\partial x_2} \frac{\partial w}{\partial x_2} \right) d\mathbf{x}.$$

On note que la forme L est linéaire en les deux variables indépendantes, si bien que (5.18) se met sous forme de système d'équations algébriques linéaires par rapport aux coefficients $\alpha_{\mathbf{y}}$

$$\sum_{\mathbf{y} \in \Omega_h} \alpha_{\mathbf{y}} L(\varphi_{\mathbf{y}}, \varphi_{\mathbf{z}}) = (f, \varphi_{\mathbf{z}}), \qquad \mathbf{z} \in \Omega_h. \tag{5.19}$$

Les inconnues sont en nombre $(2N_1 - 1)(N_2 - 1)$, et le système a autant d'équations que d'inconnues.

En calcul automatique, il y a intérêt à écrire (5.19) en notations matricielles. On numérote les nœuds de Ω_h dans l'ordre suivant: $(h_1, h_2), (h_1, 2h_2), \ldots, (h_1, (N_2-1)h_2), (2h_1, h_2), \ldots, ((2N_1-1)h_1, (N_2-1)h_2)$. Cela détermine un numérotage unique des équations et des inconnues du système concerné dont la matrice s'écrit sous forme (bloc-tridiagonale)

$$\begin{bmatrix} B_1 & A_2^T & & & 0 \\ A_2 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & A_{2N_1-1}^T \\ 0 & & & A_{2N_1-1} & B_{2N_1-1} \end{bmatrix}.$$

A_i, B_i sont des matrices carrées d'ordre $(N_2-1)^2$, les matrices B_i étant tridiagonales et A_i diagonales:

$$B_i = \begin{cases} a_1\left(\dfrac{h_1}{h_2} D + \dfrac{2h_2}{h_1} I\right), & i = 1, 2, \ldots, N_1 - 1, \\ \dfrac{a_1 + a_2}{2}\left(\dfrac{h_1}{h_2} D + \dfrac{2h_2}{h_1} I\right), & i = {}_1N, \\ a_2\left(\dfrac{h_1}{h_2} D + \dfrac{2h_2}{h_1} I\right), & i = N_1 + 1, \ldots, 2N_1 - 1, \end{cases}$$

$$A_i = \begin{cases} -a_1 \dfrac{h_2}{h_1} I, & i = 2, 3, \ldots, N_1, \\ -a_2 \dfrac{h_2}{h_1} I, & i = N_1 + 1, \ldots, 2N_1 - 1. \end{cases}$$

Ici I est une matrice unité et D une matrice tridiagonale de dimension $(N_2-1)^2$:

$$I = \begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & \cdot & & \\ & & & \cdot & \\ 0 & & & 1 & \\ & & & & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & -1 \\ 0 & & & -1 & 2 \end{bmatrix}.$$

On voit que la matrice du système (5.19) est irréductiblement diagonalement dominante (voir [65]), donc non dégénérée. Aussi le système (5.19) n'a pas d'autre solution que $\{\alpha_y\}$ qui définit u^h de façon unique.

Quelle est la précision de la solution approchée obtenue? On introduit l'interpolant linéaire par morceaux $\tilde{u}^h \in H^h$ de u

$$\tilde{u}^h(\mathbf{x}) = \sum_{\mathbf{y} \in \Omega_h} u(\mathbf{y})\, \varphi_{\mathbf{y}}(\mathbf{x}).$$

On note que les propriétés des fonctions φ_y permettent de dire que $\tilde{u}^h(\mathbf{x}) = u(\mathbf{x})\ \forall\, \mathbf{x} \in \bar{\Omega}_h$, ce qui justifie le nom d'interpolant donné à $\tilde{u}^h$.

LEMME 5.1. *Les fonctions u^h et $\tilde{u}^h$ ainsi obtenues vérifient l'identité*

$$L(u^h - u,\ u^h - u) = L(u^h - u,\ \tilde{u}^h - u),$$

avec u la solution du problème (5.6).

DÉMONSTRATION. L'identité intégrale (5.6) entraîne

$$L(u,\ \varphi_{\mathbf{z}}) = (f,\ \varphi_{\mathbf{z}}), \qquad \mathbf{z} \in \Omega_h. \tag{5.20}$$

On additionne ces identités munies des poids α_z, il vient l'égalité

$$L(u,\ u^h) = (f,\ u^h) \tag{5.21}$$

car le produit scalaire et la forme L sont linéaires par rapport à la seconde variable. Si l'on opère avec (5.18), alors on a au lieu de (5.21):

$$L(u^h,\ u^h) = (f,\ u^h).$$

On fait la différence de cette égalité et de (5.21):

$$L(u^h - u,\ u^h) = 0.$$

On démontre de même que

$$L(u^h - u,\ \tilde{u}^h) = 0,$$

d'où

$$L(u^h - u,\ u^h) = L(u^h - u,\ \tilde{u}^h).$$

On obtient le résultat voulu en retranchant $L(u^h - u,\ u)$ de deux membres de cette égalité.

Nous allons déduire une inégalité simple reliant la précision de la solution approchée u^h et celle de l'interpolant $\tilde{u}^h$.

LEMME 5.2. *Les fonctions u^h et $\tilde{u}^h$ satisfont à l'inégalité*

$$|u^h - u| \leqslant |\tilde{u}^h - u| \frac{\max\{a_1, a_2\}}{\min\{a_1, a_2\}} {}^{*}.$$

DÉMONSTRATION. On aura besoin de deux inégalités auxiliaires. On applique à $L(v, w)$ l'inégalité de Cauchy-Bouniakovski, il vient

$L(v, w) \leqslant$

$$\leqslant \left(\int_\Omega a\left(\left(\frac{\partial v}{\partial x_1}\right)^2 + \left(\frac{\partial v}{\partial x_2}\right)^2 dx\right)\right)^{1/2} \left(\int_\Omega a\left(\left(\frac{\partial w}{\partial x_1}\right)^2 + \left(\frac{\partial w}{\partial x_2}\right)^2\right) dx\right)^{1/2}.$$

On remplace $a(x)$ sous le signe $\int$ par $\max\{a_1, a_2\}$ et on utilise la définition des normes $|v|$ et $|w|$:

$$L(v, w) \leqslant \max\{a_1, a_2\} |v| |w|. \tag{5.22}$$

On cherche une minoration de $L(v, v)$. Par définition d'une forme, on a

$$L(v, v) = \int_\Omega a\left(\left(\frac{\partial v}{\partial x_1}\right)^2 + \left(\frac{\partial v}{\partial x_2}\right)^2\right) dx.$$

On remplace $a(x)$ par $\min\{a_1, a_2\}$ et on utilise la définition de $|v|$, il vient

$$L(v, v) \geqslant \min\{a_1, a_2\} |v|^2. \tag{5.23}$$

On évalue les deux membres de l'identité du lemme 5.1 par les inégalités (5.22) et (5.23) où l'on pose $v = u^h - u$ et $w = \tilde{u}^h - u$:

$$\min\{a_1, a_2\} |u^h - u|^2 \leqslant L(u^h - u, u^h - u) =$$
$$= L(u^h - u, \tilde{u}^h - u) \leqslant \max\{a_1, a_2\} |u^h - u| |\tilde{u}^h - u|. \tag{5.24}$$

Si $|u^h - u| = 0$, le lemme se trouve démontré. Soit $|u^h - u| \neq 0$. On divise (5.24) par $\min(a_1, a_2) |u^h - u|$, ce qui fournit l'estimation désirée.

On apprécie l'erreur sur la solution u^h au moyen de l'erreur sur $\tilde{u}^h$.

THÉORÈME 5.3. *On suppose qu'on est, pour le problème* (5.6), *dans les conditions* (5.5). *La solution approchée u^h obtenue par la méthode de Galerkin* (5.17), (5.18) *admet l'estimation*

$$|u^h - u| \leqslant c_5 h (\|u\|_{W_2^2(\Omega_1)} + \|u\|_{W_2^2(\Omega_2)}), \tag{5.25}$$

* $|u|$ désignera jusqu'à la fin du paragraphe une des normes de la fonction u (voir $|u|$, p. 188).

où $h = \max\{h_1, h_2\}$.

DÉMONSTRATION. On évalue $|\tilde{u}^h - u|$ moyennant les inégalités de [37]

$$\int_{\Omega_i} \left(\frac{\partial \tilde{u}^h}{\partial x_j} - \frac{\partial u}{\partial x_j}\right)^2 dx \leqslant 4h^2 \| u \|^2_{W_2^2(\Omega_i)}, \qquad j = 1, 2,$$

avec $i = 1, 2$. On en fait la somme pour les deux valeurs de i, il vient

$$|\tilde{u}^h - u|^2 \leqslant 8h^2 (\| u \|^2_{W_2^2(\Omega_1)} + \| u \|^2_{W_2^2(\Omega_2)}),$$

auquel cas le résultat du lemme 5.2 entraîne

$$|u^h - u| \leqslant 2\sqrt{2}h \frac{\max\{a_1, a_2\}}{\min\{a_1, a_2\}} (\| u \|^2_{W_2^2(\Omega_1)} + \| u \|^2_{W_2^2(\Omega_2)})^{1/2}.$$

Les quantités $\| u \|_{W_2^2(\Omega_i)}$ étant finies, on a (5.25), c.q.f.d.

Ainsi, la précision de la solution u^h de la méthode de Galerkin est obtenue pour la norme $|u^h - u|$. Avec le théorème de l'immersion de $\overset{\circ}{W}{}_2^1(\Omega)$ dans $L_2(\Omega)$ (voir [26]), on évalue la même erreur en norme $L_2(\Omega)$. En effet, le théorème implique

$$\| u^h - u \|_{L_2(\Omega)} \leqslant c_6 |u^h - u|.$$

On utilise l'estimation (5.25), il vient

$$\| u^h - u \|_{L_2(\Omega)} \leqslant c_7 h.$$

Mais cette inégalité n'est pas caractéristique de la méthode de Galerkin car l'erreur en norme $L_2(\Omega)$ est en réalité une quantité du second ordre en h. On le démontre par le procédé décrit dans [39], [59], [118].

Soit le problème (5.6) avec un autre second membre. On cherche une solution w telle qu'elle vérifie l'identité intégrale

$$L(w, v) = (u^h - u, v) \qquad \forall v \in \overset{\circ}{W}{}_2^1(\Omega). \tag{5.26}$$

Comme $(u^h - u) \in L_2(\Omega)$, on a (voir le début du paragraphe) $w \in W_2^2(\Omega_1) \cap W_2^2(\Omega_2)$ et la majoration

$$\| w \|_{W_2^2(\Omega_1)} + \| w \|_{W_2^2(\Omega_2)} \leqslant c_4 \| u^h - u \|_{L_2(\Omega)}. \tag{5.27}$$

On trouve par la méthode de Galerkin (5.17), (5.18) une solution approchée $w^h \in H^h$ du problème. Cette solution satisfait également au théorème 5.3, si bien que

$$|w^h - w| \leqslant c_5 h (\| w \|_{W_2^2(\Omega_1)} + \| w \|_{W_2^2(\Omega_2)}).$$

d'où, par l'estimation (5.27),

$$|w^h - w| \leqslant c_4 c_5 h \| u^h - u \|_{L_2(\Omega)} . \tag{5.28}$$

On pose $v = u^h - u$ dans (5.26) et on utilise la propriété de commutativité des variables de la forme bilinéaire L, il vient

$$\| u^h - u \|^2_{L_2(\Omega)} = L(w, u^h - u) = L(u^h - u, w). \tag{5.29}$$

Comme $w^h \in H^h$, on a

$$L(u^h - u, w^h) = 0.$$

Avec cette égalité, (5.29) devient

$$\| u^h - u \|^2_{L_2(\Omega)} = L(u^h - u, w - w^h).$$

On applique successivement (5.22), (5.25) et (5.28):

$$\| u^h - u \|^2_{L_2(\Omega)} \leqslant \max \{a_1, a_2\} | u^h - u | | w - w^h | \leqslant$$

$$\leqslant \max \{a_1, a_2\} c_4 c_5^2 h^2 \| u^h - u \|_{L_2(\Omega)} (\| u \|_{H_2^2(\Omega_1)} + \| u \|_{H_2^2(\Omega_2)}).$$

Si $\| u^h - u \|_{L_2(\Omega)} \neq 0$, on divise l'inégalité obtenue par cette quantité et on se sert de l'estimation (5.15), il vient

$$\| u^h - u \|_{L_2(\Omega)} \leqslant c_8 h^2 \| f \|_{L_2(\Omega)} . \tag{5.30}$$

Ce résultat est juste à fortiori pour $\| u^h - u \|_{L_2(\Omega)} = 0$.

Ainsi, l'erreur commise sur la solution approchée u^h du problème (5.1) à (5.5) résolu par la méthode de Galerkin (5.17), (5.18) est une quantité de l'ordre de h^2 pour la norme $L_2(\Omega)$.

4.6. Dégagement des singularités

Dans ce paragraphe, nous décrirons un procédé de dégagement des singularités qui apparaissent au voisinage des points anguleux. S'agissant des problèmes en dimension deux, on utilise volontiers, en plus du resserrement du réseau autour du point anguleux, le procédé qui consiste à ajouter aux fonctions de base locales usuelles des méthodes de Ritz et de Boubnov-Galerkin (ce sont par exemple les fonctions linéaires par morceaux sur les triangles qui portent le nom de Courant; voir [81]) des fonctions singulières qui décrivent bien le comportement de la partie irrégulière de la solution au voisinage du point anguleux (voir [38], [60], [64], [132]). Ce sont en général les termes principaux des développements asymptotiques de la forme (4.7) qu'on transforme de façon qu'ils vérifient les conditions aux limites essentielles. On prend des fois des termes suivants pour

atteindre une précision en h^2. On note que ces termes doivent être connus à des constantes multiplicatives près qui deviennent les inconnues auxiliaires du système algébrique de dimension finie intervenant dans les méthodes de Ritz et de Boubnov-Galerkin. Dans deux paragraphes précédents, on a motivé les résultats moyennant un des procédés possibles de dégagement des termes principaux des développements asymptotiques. On l'utilise constructivement pour former les fonctions de base singulières de la méthode décrite.

Voici une autre façon de traiter les singularités au voisinage des points anguleux. On remplace le problème initial par plusieurs problèmes dont l'un est à données concordantes et possède une solution régulière, et les autres ont des données n'ayant pas cette propriété et un domaine de définition plus simple qui se prête bien au calcul numérique et où l'on introduit commodément les coordonnées polaires. En dimension deux, ce domaine standard peut être un secteur circulaire d'angle d'ouverture déterminé. Les problèmes sont résolus simultanément par le procédé alterné de Schwarz. La méthode décrite s'est formée sous l'influence de [12], où l'on justifie la résolution de l'équation de Laplace sur un polygone par la technique de Schwarz pour les réseaux de discrétisation polaires et rectangulaires.

Désireux de simplifier l'exposé, nous nous bornons à un domaine avec un seul point anguleux.

Soit Ω un domaine lipschitzien borné de frontière Γ régulière partout sauf au point $(0, 0)$ au voisinage duquel Γ est formée de deux segments de droite qui se coupent sous l'angle $\Phi \in (0, 2\pi)$ (fig. 4.9). Soit, dans Ω, l'équation

$$-\Delta u = f \tag{6.1}$$

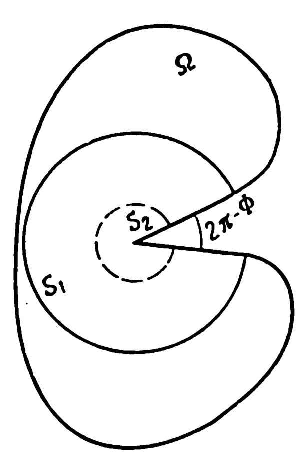


Fig. 4.9. Domaine avec point anguleux

avec la condition aux limites

$$u = g \quad \text{sur} \quad \Gamma. \tag{6.2}$$

On suppose que

$$f \in C^{2+\alpha}(\bar{\Omega}) \tag{6.3}$$

avec une constante $\alpha \in (0, 1)$. On désigne par Γ_1 le tronçon de Γ qui complète les segments de droite issus du point $(0, 0)$ de façon à obtenir Γ tout entière. Γ_1 est supposé suffisamment régulier:

$$\Gamma_1 \in C^{4+\alpha}. \tag{6.4}$$

On exige que

$$g \in C^{4+\alpha}(\Gamma \setminus (0,0)) \tag{6.5}$$

et que la fonction g admet en (0, 0) des limites finies, ainsi que ses dérivées jusqu'à l'ordre 4 inclus prises le long de Γ. On connaît pour les angles π/j, $j = 2, 3, \ldots$, des critères simples de concordance des données qui garantissent la régularité de la solution du problème de Dirichlet (voir [51]). Les autres critères ont un caractère intégral compliqué, si bien qu'on n'assujettit dans ce paragraphe le second membre f et la fonction g à aucune condition de concordance. On note que la solution du problème (6.1), (6.2) peut même ne pas être continue sur $\bar{\Omega}$, mais qu'elle jouit de la propriété

$$u \in C^{4+\alpha}(\bar{\Omega}_1) \tag{6.6}$$

quel que soit $\Omega_1 \subset \Omega$ dont la distance à (0, 0) est une quantité positive (voir [26]).

On introduit les coordonnées polaires (r, φ) de centre (0, 0). On choisit un nombre $a > 0$ tel que l'intersection de Ω et du disque de rayon a centré en (0, 0) soit le secteur

$$S_1 = \{(r, \varphi);\ 0 < r < a,\ 0 < \varphi < \Phi\}.$$

Le secteur circulaire de rayon $a/6$ sera noté S_2:

$$S_2 = \{(r, \varphi);\ 0 < r < a/6,\ 0 < \varphi < \Phi\}.$$

On verra plus loin qu'il y a intérêt à prendre a le plus grand possible.

On suppose que le domaine Ω est inclus dans le carré $\{(x, y);$ $-b < x < b,\ -b < y < b\}$. On couvre celui-ci d'un réseau carré de pas $h = b/N$ formé par les lignes $x_i = ih$, $y_j = jh$, $i, j = -N, \ldots, N$. Soit Ω_h l'ensemble des nœuds intérieurs du réseau et Γ_h l'ensemble des nœuds frontières (voir n° 4.2.1). On suppose de plus que

$$D_h = \Omega_h \setminus S_2.$$

On définit pour D_h l'ensemble ∂D_h des nœuds frontières. Il correspond à chaque point de D_h dans les directions parallèles aux axes de coordonnées quatre nœuds voisins, i.e. quatre nœuds le plus proches qui sont, ou bien dans Γ_h, ou bien dans Ω_h. On appelle $\bar{D}_h$ la réunion de l'ensemble des points de D_h et de l'ensemble des points voisins. On pose naturellement

$$\partial D_h = \bar{D}_h \setminus D_h.$$

On observe que la distance de chaque nœud de $\bar{D}_h$ et du point (0, 0) est au moins égale à $a/6 - h$. Puisqu'on s'intéresse à la façon dont la solution approchée se comporte quand $h \to 0$, on estimera dans la suite que h est suffisamment petit, par exemple

$$a/6 - h \geq a/8.$$

On définit en chaque $(x, y) \in D_h$ l'opérateur aux différences à cinq points

$$L_h v(x, y) \equiv$$
$$\equiv \frac{1}{h}\left[\frac{1}{\delta_1} v(x+\delta_1, y) + \frac{1}{\delta_2} v(x-\delta_2, y) + \frac{1}{\theta_1} v(x, y+\theta_1) + \right.$$
$$\left. + \frac{1}{\theta_2} v(x, y-\theta_2) - \left(\frac{1}{\delta_1} + \frac{1}{\delta_2} + \frac{1}{\theta_1} + \frac{1}{\theta_2}\right) v(x, y)\right]. \quad (6.7)$$

où $(x+\delta_1, y)$, $(x-\delta_2, y)$, $(x, y+\theta_1)$, $(x, y-\theta_2)$ sont les points de $\overline{D}_h$ voisins du nœud $(x, y) \in D_h$. Si les quatre distances δ_1, δ_2, θ_1, θ_2 sont h, alors le nœud (x, y) est dit régulier. Tous les nœuds réguliers de D_h forment un ensemble noté D_h^r. Les nœuds restants de D_h s'appellent nœuds irréguliers. On pose $D_h^{ir} = D_h \diagdown D_h^r$.

Comme tous les nœuds de $\overline{D}_h$ sont distants de $(0, 0)$ d'une quantité au moins égale à $a/8$, l'opérateur aux différences L_h sur u est approché à l'ordre 2 aux nœuds réguliers et à l'ordre 0 aux nœuds restants. En effet, on vérifie sans peine que

$$L_h u(x, y) - \Delta u(x, y) = \alpha^h(x, y), \qquad (x, y) \in D_h,$$

et

$$|\alpha^h(x, y)| \leqslant h^2 c_1 \qquad \forall (x, y) \in D_h^r, \quad (6.8)$$

$$|\alpha^h(x, y)| \leqslant c_2 \qquad \forall (x, y) \in D_h^{ir}, \quad (6.9)$$

où les constantes

$$c_1 = \frac{1}{12}\left(\max_{\overline{\Omega}_2}\left|\frac{\partial^4 u}{\partial x^4}\right| + \max_{\overline{\Omega}_2}\left|\frac{\partial^4 u}{\partial y^4}\right|\right), \quad (6.10)$$

$$c_2 = 2\left(\max_{\overline{\Omega}_2}\left|\frac{\partial^2 u}{\partial x^2}\right| + \max_{\overline{\Omega}^2}\left|\frac{\partial^2 x}{\partial y^2}\right|\right). \quad (6.11)$$

Ici $\overline{\Omega}_2$ est l'ensemble des points de $\overline{\Omega}$ dont la distance à $(0, 0)$ est au moins égale à $a/8$.

Soit le problème discrétisé

$$-L_h v^h = f \quad \text{sur} \quad D_h, \quad (6.12)$$

$$v^h = g \quad \text{sur} \quad \partial D_h \cap \Gamma_h, \quad (6.13)$$

$$v^h = w^h \quad \text{sur} \quad \partial D_h \cap \Omega_h, \quad (6.14)$$

avec w^h une fonction discrète arbitraire définie sur $\partial D_h \cap \Omega_h$. L'opérateur L_h vérifie le principe du maximum (voir [42]), si bien que le problème homogène associé au système algébrique linéaire (6.12) à (6.14) admet la seule solution triviale. Aussi le problème non homogène (6.12) à (6.14) possède pour tout second membre une solution unique v^h.

LEMME 6.1. *La solution du problème* (6.12) *à* (6.14) *admet l'estimation*

$$\max_{\bar{D}_h} |v^h - u| \leq c_3 h^2 + \max_{\partial D_h \cap \Omega_h} |w^h - u|.$$

DÉMONSTRATION. On représente la solution v^h par la somme $v_1^h + v_2^h$, les fonctions v_i^h étant les solutions des problèmes

$$\begin{aligned} -L_h\, v_1^h &= 0 && \text{sur } D_h, \\ v_1^h &= 0 && \text{sur } \partial D_h \cap \Gamma_h, \\ v_1^h &= w^h - u && \text{sur } \partial D_h \cap \Omega_h; \end{aligned} \tag{6.15}$$

$$\begin{aligned} -L_h\, v_2^h &= f && \text{sur } D_h, \\ v_2^h &= g && \text{sur } \partial D_h \cap \Gamma_h \\ v_2^h &= u && \text{sur } \partial D_h \cap \Omega_h. \end{aligned} \tag{6.16}$$

Le premier problème satisfait au principe du maximum qui implique

$$\max_{\bar{D}_h} |v_1^h| \leq |\max_{\partial D_h \cap \Omega_h} |w^h - u||. \tag{6.17}$$

La solution exacte u vérifie les égalités

$$\begin{aligned} -L_h\, u &= f + \alpha^h && \text{sur } D_h, \\ u &= g && \text{sur } \partial D_h \cap \Gamma_h. \end{aligned}$$

La différence $\varepsilon^h = u - v_2^h$ est donc solution de

$$\begin{aligned} -L_h\, \varepsilon^h &= \alpha^h && \text{sur } D_h, \\ \varepsilon^h &= 0 && \text{sur } \partial D_h. \end{aligned} \tag{6.18}$$

avec α^h évalués par (6.8), (6.9).

Soit, sur $\bar{D}_h$, la fonction

$$\rho^h(x, y) = \frac{c_1 h^2}{4}(2b^2 - x^2 - y^2) + \begin{cases} c_2 h^2 & \text{si } (x, y) \in D_h, \\ 0 & \text{si } (x, y) \in \partial D_h. \end{cases}$$

Il est immédiat de vérifier que

$$\rho^h(x, y) \geq 0 \qquad \forall (x, y) \in \partial D_h,$$

$$-L_h\, \rho^h(x, y) \geq \begin{cases} c_1 h^2 & \text{si } (x, y) \in D_h^r, \\ c_2 & \text{si } (x, y) \in D_h^{ir}. \end{cases}$$

Vu les estimations (6.8), (6.9), on aboutit au résultat

$$-L_h\,\rho^h(x, y) \geqslant |\alpha^h(x, y)| \qquad \forall (x, y) \in D_h.$$

On a donc par un théorème de comparaison (voir [42])

$$|\varepsilon^h(x, y)| \leqslant \rho^h(x, y) \qquad \forall (x, y) \in \overline{D}_h.$$

Avec la formule donnant la fonction ρ^h, on en tire

$$\max_{\overline{D}_h} |\varepsilon^h| \leqslant \frac{b^2 c_1 h^2}{2} + c_2 h^2.$$

On pose

$$c_3 = c_2 + \frac{b^2}{2} c_1,$$

et on obtient

$$\max_{\overline{D}_h} |v^h - u| \leqslant \max_{\overline{D}_h} |v_2^h - u| + \max_{\overline{D}_h} |v_1^h| \leqslant c_3 h^2 + \max_{\partial D_h \cap \Omega_h} |w^h - u|.$$

Le lemme se trouve démontré.

Soit, dans le secteur S_1, le problème

$$\begin{aligned} -\Delta w &= f \quad \text{dans} \quad S_1, \\ w &= g \quad \text{sur} \quad \partial S_1 \cap \Gamma, \\ w &= u \quad \text{sur} \quad \partial S_1 \setminus \Gamma. \end{aligned} \tag{6.19}$$

Il admet évidemment pour solution la fonction u. En passant aux coordonnées polaires, le problème se récrit *

$$-\frac{\partial^2 w}{\partial r^2} - \frac{1}{r}\frac{\partial w}{\partial r} - \frac{1}{r^2}\frac{\partial^2 w}{\partial \varphi^2} = f, \tag{6.20}$$

$$r \in (0, a), \qquad \varphi \in (0, \Phi),$$

$$\left.\begin{aligned} w(r, 0) &= g(r, 0), \\ w(r, \Phi) &= g(r, \Phi). \end{aligned}\right\} \quad r \in (0, a), \tag{6.21}$$

$$w(a, \varphi) = u(a, \varphi), \qquad \varphi \in (0, \Phi). \tag{6.22}$$

Si les conditions aux limites (6.21) ne sont pas homogènes, la fonction inconnue est remplacée par une autre, à savoir

$$w_1(r, \varphi) = w(r, \varphi) - \left(1 - \frac{\varphi}{\Phi}\right) g(r, 0) - \frac{\varphi}{\Phi} g(r, \Phi). \tag{6.23}$$

---

* On conserve la notation des fonctions dans lesquelles on a effectué le changement de coordonnées.

On note que la régularité de w_1 par rapport à φ coïncide avec celle de w. Le second membre de (6.20) et la condition aux limites (6.22) deviennent par ce changement d'inconnue

$$-\frac{\partial^2 w_1}{\partial r^2}-\frac{1}{r}\frac{\partial w_1}{\partial r}-\frac{1}{r^2}\frac{\partial^2 w_1}{\partial \varphi^2}=f_1, \tag{6.24}$$

$$r\in(0,\ a),\qquad \varphi\in(0,\ \Phi),$$

$$w_1(r,\ 0)=w_1(r,\ \Phi)=0,\qquad r\in(0,\ a), \tag{6.25}$$

$$w_1(a,\ \varphi)=z_1(a,\ \varphi),\qquad \varphi\in(0,\ \Phi), \tag{6.26}$$

où

$$z_1(a,\ \varphi)=u(a,\ \varphi)-\left(1-\frac{\varphi}{\Phi}\right)g(a,\ 0)-\frac{\varphi}{\Phi}g(a,\ \Phi), \tag{6.27}$$

$$f_1(r,\ \varphi)=f(r,\ \varphi)+\left(1-\frac{\varphi}{\Phi}\right)\left[\frac{\partial^2 g}{\partial r^2}(r,\ 0)+\frac{1}{r}\frac{\partial g}{\partial r}(r,\ 0)\right]+$$
$$+\frac{\varphi}{\Phi}\left[\frac{\partial^2 g}{\partial r^2}(r,\Phi)+\frac{1}{r}\frac{\partial g}{\partial r}(r,\Phi)\right]. \tag{6.28}$$

Le problème (6.24) à (6.26) est résolu par la méthode de la séparation des variables. Soit $F_k(r)$ les coefficients de Fourier de $f_1(r,\ \varphi)$:

$$f_1(r,\varphi)=\sum_{k=1}^{\infty}F_k(r)\sin k\lambda\varphi,$$

où

$$\lambda=\frac{\pi}{\Phi}, \tag{6.29}$$

les coefficients étant calculés par les formules

$$F_k(r)=\frac{2}{\Phi}\int_0^{\Phi}f_1(r,\ \varphi)\sin k\lambda\varphi\ d\varphi. \tag{6.30}$$

On développe la solution w_1 du problème (6.24) à (6.26) en série de Fourier:

$$w_1(r,\varphi)=\sum_{k=1}^{\infty}W_k(r)\sin k\lambda\varphi. \tag{6.31}$$

Les coefficients W_k sont solutions des équations

$$-W_k''-\frac{1}{r}W_k'+\frac{k^2\lambda^2}{r^2}W_k=F_k,\qquad r\in(0,\ a). \tag{6.32}$$

avec les conditions aux limites

$$|W_k(0)| < \infty, \qquad W_k(a) = z_k. \tag{6.33}$$

où

$$z_k = \frac{2}{\Phi}\int_0^{\Phi} z(a, \varphi) \sin k\lambda\varphi \, d\varphi. \tag{6.34}$$

Ce problème aux limites possède pour solution

$$W_k(r) = \frac{-r^{\lambda k}}{2\lambda k a^{2\lambda k}}\int_0^a F_k(\rho)\,\rho^{\lambda k+1}\,d\rho + \frac{1}{2\lambda k} r^{\lambda k}\int_r^a F_k(\rho)\,\rho^{1-\lambda k}\,d\rho +$$

$$+ \frac{1}{2\lambda k} r^{-\lambda k}\int_0^r F_k(\rho)\,\rho^{\lambda k+1}\,d\rho + \frac{r^{\lambda k}}{a^{\lambda k}} z_k. \tag{6.35}$$

Les M premiers coefficients seuls nous intéressent car on peut prendre pour solution approchée du problème (6.24) à (6.26) la série de Fourier tronquée

$$w_2^h(r, \varphi) = \sum_{k=1}^{M} W_k(r) \sin k\lambda\varphi. \tag{6.36}$$

Comment l'écart entre w_2^h et la solution w_1 dépend du nombre M de termes du développement (6.36)? La propriété d'orthogonalité des fonctions sin $k\lambda\varphi$ pour divers k entraîne

$$W_k(r) = \frac{2}{\Phi}\int_0^{\Phi} w_1(r, \varphi) \sin k\lambda\varphi \, d\varphi. \tag{6.37}$$

On signale une grande « réserve de dérivabilité » par rapport à φ de la fonction w_1. On s'intéresse surtout aux valeurs de W_k et w_1 dépendant de $r \in [a/8, a/6]$, i.e. de r appartenant au segment où w_1 admet des dérivées bornées continues de quatre premiers ordres. Dans ce cas, on garantit (voir [49]) la décroissance suivante des coefficients:

$$W_k(r)| \leqslant d_1/k^3, \qquad k = 1, 2, \ldots, \qquad \forall\, r \in [a,/8\ a/6]. \tag{6.38}$$

Cette inégalité permet d'évaluer le reste de la série de Fourier de w_1. On a

$$|w_1(r, \varphi) - w_2^h(r, \varphi)| = \left|\sum_{k=M+1}^{\infty} W_k(r) \sin k\lambda\varphi\right| \leqslant$$

$$\leqslant \sum_{k=M+1}^{\infty} |W_k(r)| \leqslant d_1 \sum_{k=M+1}^{\infty} \frac{1}{k^3} < \frac{d_1}{2M^2} \qquad \forall\, r \in [a/8, a/6]. \tag{6.39}$$

On passe au dernier membre de cette inégalité à l'aide de l'estimation

$$\sum_{k=M+1}^{\infty} \frac{1}{k^{p+1}} \leq \frac{1}{p} \frac{1}{M^p}, \quad p > 0 \tag{6.40}$$

(voir [48]). Pour qu'il y ait compatibilité entre la précision de la série de Fourier tronquée et celle du schéma aux différences, on pose M du même ordre de grandeur que N. On admet, pour simplifier l'exposé, que $M = N$. Comme $h = b/N$, l'inégalité (6.39) devient

$$| w_1 (r, \varphi) - w_2^h (r, \varphi) | \leqslant \frac{d_1}{2 b^2} h^2. \tag{6.41}$$

La recherche de w_2^h relève du calcul des intégrales auxquelles on substitue les formules de quadrature. On procède comme suit. On remplace f_1 à valeurs données aux nœuds (ρ_i, ψ_j), où $\rho_i = i\frac{a}{N}$, $\psi_j = j\frac{\Phi}{N}$, $i, j = 0, 1, \ldots, N$, par son prolongement $\tilde{f}_1$ linéaire par morceaux par rapport à chaque variable (voir [37]), si bien qu'on a à l'intérieur de chaque rectangle élémentaire une fonction linéaire en ρ et ψ:

$$\begin{aligned}\tilde{f}_1 (\rho, \psi) = \frac{N^2}{a\Phi} \{(\rho_{i+1} - \rho)[(\psi_{j+1} - \psi) f_1 (\rho_i, \psi_j) + (\psi - \psi_j) \times \\ \times f_1 (\rho_i, \psi_{i+1})] + (\rho - \rho_i)[\psi_{j+1} - \psi) f_1 (\rho_{i+1}, \psi_j) + \\ + (\psi - \psi_j) (\rho_{i+1}, \psi_{j+1})]\}.\end{aligned} \tag{6.42}$$

si $\psi \in [\psi_j, \psi_{j+1}], \rho \in [\rho_i, \rho_{i+1}]$.

On substitue à f_1 de (6.30) son prolongement $\tilde{f}_1$, il vient les fonctions linéaires par morceaux

$$\tilde{F}_k (\rho) = \frac{2}{\Phi} \int_0^{\Phi} \tilde{f}_1 (\rho, \psi) \sin k\lambda\psi \, d\psi \tag{6.43}$$

qui sont bien définies par leurs valeurs aux points ρ_i parce que $\tilde{F}_k (\rho) = \frac{N}{a} [(\rho_{i+1} - \rho) \tilde{F}_k (\rho_i) + (\rho - \rho_i) \tilde{F}_k (\rho_{i+1})]$ si $\rho \in [\rho_i, \rho_{i+1}]$. Aussi on calcule les intégrales (6.43) pour $\rho = \rho_j$, $j = 0, 1, \ldots, N$, seuls.

On remplace de même la fonction z de (6.34) par son prolongement linéaire par morceaux

$$\tilde{z}(\psi) = \frac{N}{\Phi} [(\psi_{j+1} - \psi) z (\psi_j) + (\psi - \psi_j) z (\psi_{j+1})]. \tag{6.44}$$

auquel cas

$$\tilde{z}_k = \frac{2}{\Phi}\int_0^\Phi \tilde{z}(\psi)\sin k\lambda\psi\, d\psi \tag{6.45}$$

constitue (comme (6.42)) la formule de quadrature des trapèzes avec poids $\sin k\lambda\psi$.

On calcule de façon approchée les coefficients de Fourier:

$$\tilde{W}_k(r) = \frac{-r^{\lambda k}}{2\lambda k a^{2\lambda k}}\int_0^a \tilde{F}_k(\rho)\,\rho^{\lambda k+1}\,d\rho + \frac{r^{\lambda k}}{2\lambda k}\int_r^a \tilde{F}_k(\rho)\,\rho^{1-\lambda k}\,d\rho +$$

$$+\frac{r^{-\lambda k}}{2\lambda k}\int_0^r \tilde{F}_k(\rho)\,\rho^{\lambda k+1}\,d\rho + \frac{r^{\lambda k}}{a^{\lambda k}}\tilde{z}_k. \tag{6.46}$$

Avec ces coefficients, on a, au lieu de w_2^h,

$$w_3^h(r, \varphi) = \sum_{k=1}^N \tilde{W}_k(r)\sin k\lambda\varphi. \tag{6.47}$$

On se propose d'évaluer la différence de w_2^h et w_3^h.

Selon les auteurs [37], [67], les interpolations linéaire et bilinéaire sont exactes à l'ordre deux pour les fonctions ayant des dérivées secondes bornées. Aussi

$$|f_1(\rho, \psi) - \tilde{f}_1(\rho, \psi)| \leqslant d_2 h^2 \qquad \forall(\rho, \psi) \in [0, a] \times [0, \Phi],$$

$$|z(\psi) - \tilde{z}(\psi)| \leqslant d_3 h^2 \qquad \forall\psi \in [0, \Phi].$$

Il résulte de (6.34) et (6.45)

$$|z_k - \tilde{z}_k| \leqslant \frac{2d_3h^2}{\Phi}\int_0^\Phi |\sin k\lambda\psi|\, d\psi \leqslant 2\, d_3 h^2, \qquad k = 1, 2, \ldots$$

De même,

$$|F_k(\rho) - \tilde{F}_k(\rho)| \leqslant 2d_2h^2, \qquad \rho \in [0, a], \qquad k = 1, 2, \ldots$$

On retranche (6.46) de (6.35) et on utilise les estimations obtenues, il vient

$$|W_k(r) - \tilde{W}_k(r)| \leqslant \frac{d_2h^2}{2\lambda k}\left(\frac{r^{\lambda k}}{a^{2\lambda k}}\int_0^a \rho^{\lambda k+1}\,d\rho + r^{\lambda k}\int_r^a \rho^{1-\lambda k}\,d\rho + \right.$$

$$\left. + r^{-\lambda k}\int_0^r \rho^{\lambda k+1}\,d\rho\right) + \frac{r^{\lambda k}}{a^{\lambda k}}\, 2\, d_3 h^2.$$

On fait l'hypotèse que $r \in [a/8, a/6]$. Le premier terme entre parenthèses admet l'estimation simple suivante:

$$\frac{r^{\lambda k}}{a^{2\lambda k}} \int_0^a \rho^{\lambda k+1}\, d\rho = \frac{r^{\lambda k} a^{2-\lambda k}}{\lambda k + 2} \leqslant \frac{a^2}{k\lambda 6^{\lambda k}} \leqslant \frac{a^2}{k\lambda}.$$

S'agissant de $r^{\lambda k} \int_r^a \rho^{1-\lambda k}\, d\rho$, deux cas peuvent se présenter. Si $k = 2/\lambda$, alors

$$r^{\lambda k} \int_r^a \rho^{1-\lambda k}\, d\rho = r^2 (\ln a - \ln r) \leqslant \frac{a^2}{36} \ln 8.$$

Dans le cas contraire (i.e. si $\lambda k \neq 2$), on a

$$r^{\lambda k} \int_r^a \rho^{1-\lambda k}\, d\rho = \frac{r^{\lambda k} \,|\, a^{2-\lambda k} - r^{2-\lambda k} \,|}{|\, 2 - \lambda k \,|} \leqslant \frac{a^2}{|\, 2 - \lambda k \,|}.$$

Le troisième terme est apprécié d'une manière particulièrement simple:

$$r^{-\lambda k} \int_0^r \rho^{\lambda k+1}\, d\rho = \frac{r^2}{\lambda k + 2} \leqslant \frac{v^2}{\lambda k}.$$

Ces formules montrent qu'on a pour tout $k \leqslant 4/\lambda$

$$|\, W_k(r) - \widetilde{W}_k(r) \,| \leqslant d_4 h^2.$$

Si $\lambda k > 4$, on a $|\, 2 - \lambda k \,| \geqslant \lambda k/2$, donc

$$|\, W_k(r) - \widetilde{W}_k(r) \,| \leqslant \frac{d_5 h^2}{k^2} + 2 d_3 h^2 \left(\frac{1}{6}\right)^{\lambda k}, \qquad d_5 = \frac{4 a^2 d_2}{\lambda^2}.$$

Etant données ces estimations, les formules (6.36), (6.47) entraînent

$$|\, w_2^h(r, \varphi) - w_3^h(r, \varphi) \,| \leqslant \sum_{k=1}^{N} |\, \widetilde{W}_k(r) - W_k(r) \,| \leqslant$$

$$\leqslant \sum_{k=1}^{[4/\lambda]} d_4 h^2 + \sum_{k=[4/\lambda]+1}^{N} \left(\frac{d_5 h^2}{k^2} + 2 d_3 h^2 \left(\frac{1}{6}\right)^{\lambda k} \right) \leqslant$$

$$\leqslant \frac{4 d_4}{\lambda} h^2 + d_5 h^2 \sum_{k=1}^{\infty} \frac{1}{k^2} + 2 d_3 h^2 \sum_{k=1}^{\infty} \left(\frac{1}{6}\right)^{\lambda k}.$$

La première somme du dernier membre est évaluée moyennant l'inégalité (6.40), et on somme dans la seconde la progression géométrique. Finalement,

$$|w_2^h(r, \varphi) - w_3^h(r, \varphi)| \leqslant \left(\frac{4d_4}{\lambda} + 2d_5 + \frac{2d_3 6^\lambda}{6^\lambda - 1}\right) h^2 = d_6 h^2 \quad (6.48)$$

$$\forall r \in [a/8, a/6], \qquad \forall \varphi \in [0, \Phi].$$

On fait la somme de w_3^h et du polynôme retranché plus haut, ce qui donne au lieu de w (identiquement égale à u sur S_1) la fonction

$$w_4^h(r, \varphi) = w_3^h(r, \varphi) + \left(1 - \frac{\varphi}{\Phi}\right) g(r, 0) + \frac{\varphi}{\Phi} g(r, \Phi), \quad (6.49)$$

solution approchée du problème primitif (6.20) à (6.22). On réunit les estimations (6.41) et (6.48) en l'inégalité caractéristique de la précision de w_4^h dans la bande $[a/8, a/6] \times [0, \Phi]$:

$$|w_4^h(r, \varphi) - u(r, \varphi)| \leqslant \left(d_6 + \frac{d_1}{2b^2}\right) h^2 = d_7 h^2. \quad (6.50)$$

On applique la méthode d'approximation (6.42) à (6.47) au problème

$$\begin{aligned} \Delta w_5 &= 0 \quad \text{dans} \quad S_1, \\ w_5 &= 0 \quad \text{sur} \quad \partial S_1 \cap \Gamma, \\ w_5 &= z_2 \quad \text{sur} \quad \partial S_1 \setminus \Gamma. \end{aligned} \quad (6.51)$$

On passe aux coordonnées polaires et on note qu'il suffit de définir z_2 aux points (a, φ_i), $i = 0, \ldots, N$. Les formules (6.42) à (6.47) s'écrivent pour (6.51)

$$\begin{aligned} \tilde{Z}_k &= \frac{2}{\Phi} \int_0^\Phi \tilde{z}_2(a, \psi) \sin k\lambda\psi \, d\psi, \\ \tilde{W}_k(r) &= \frac{r^{\lambda k}}{a^{\lambda k}} \tilde{Z}_k, \\ k &= 1, \ldots, N, \end{aligned} \quad (6.52)$$

avec $\tilde{z}_2$ l'interpolant linéaire par morceaux de z_2 et

$$w_6^h(r, \varphi) = \sum_{k=1}^{N} \tilde{W}_k(r) \sin k\lambda\varphi. \quad (6.53)$$

On demande la relation entre la décroissance de w_6^k en valeur absolue et la distance à la portion de frontière avec les conditions aux limites non nulles. On introduit les notations

$$\| \tilde{z}_2 \|_C = \max_{0<\varphi<\Phi} | \tilde{z}_2(a, \varphi)| = \max_{0\leq j\leq N} | z_2(a, \varphi_j)|.$$

Il résulte de (6.52)

$$| \tilde{W}_k(r) | \leq \frac{r^{\lambda k}}{a^{\lambda k}} \frac{2}{\Phi} \int_0^{\Phi} | \sin k\lambda\varphi | \, d\varphi \, \| \tilde{z}_2 \|_C \leq \frac{4}{\pi} \frac{r^{\lambda k}}{a^{\lambda k}} \| \tilde{z}_2 \|_C.$$

On examine les valeurs de $\tilde{W}_k(r)$ dans la bande $[a/8, a/6]$ sous l'hypothèse de $\lambda \geq 1/2$. Aussi

$$| \tilde{W}_k(r) | \leq \frac{4}{\pi} \left(\frac{1}{6}\right)^{k/2} \| \tilde{z}_2 \|_C \qquad \forall r \in [a/8, a/6].$$

On récrit (6.53) à la lumière de cette majoration:

$$| w_6^k(r, \varphi) | \leq \sum_{k=1}^{N} | \tilde{W}_k(r) | \, | \sin k\lambda\varphi | \, d\varphi \leq$$

$$\leq \sum_{k=1}^{\infty} \frac{4}{\pi} \left(\frac{1}{6}\right)^{k/2} \| \tilde{z}_2 \|_C = \frac{4}{\pi(\sqrt{6}-1)} \| z_2 \|_C.$$

Avec la notation

$$q = \frac{4}{\pi(\sqrt{6}-1)}, \tag{6.54}$$

on a

$$\max_{\substack{a/8\leq r\leq a/6 \\ 0\leq\varphi\leq\Phi}} | w_6^k(r, \varphi) | \leq q \| \tilde{z}_2 \|_C. \tag{6.55}$$

Soit le problème

$$\begin{aligned} \Delta w_7 &= f \quad \text{dans} \quad S_1, \\ w_7 &= g \quad \text{sur} \quad \partial S_1 \cap \Gamma, \\ w_7 &= z \quad \text{sur} \quad \partial S_1 \setminus \Gamma, \end{aligned} \tag{6.56}$$

avec z une fonction définie sur $\partial S_1 \setminus \Gamma$. La solution approchée de (6.56) est évaluée à l'aide des estimations (6.50) et (6.55).

LEMME 6.2. *On suppose que le problème* (6.56) *est résolu par la méthode* (6.42) *à* (6.47). *La solution approchée* w_8^h *dépendant de* $r \in [a/8, a/6]$ *admet la majoration*

$$|w_8^h(r, \varphi) - u(r, \varphi)| \leqslant d_7 h^2 + q \max_{0 \leqslant j \leqslant N} |z(a, \varphi_j) - u(a, \varphi_j)|, \quad (6.57)$$

où la constante d_7 *est prise dans* (6.50) *et* q *est introduit dans* (6.54).

On décrit la méthode de résolution du problème (6.1), (6.2) dans son ensemble, qui emploie le procédé alterné de Schwarz (voir [89]). On procède par itérations, chaque pas se faisant en trois temps.

ETAPE PREMIÈRE. On suppose connues les valeurs $z_{i-1}^h(a, \varphi_j)$, $j = 0, \ldots, N$ à partir du pas $i - 1$. On aborde le problème

$$\begin{aligned} \Delta v_i &= f \quad \text{dans} \quad S_1, \\ v_i &= g \quad \text{sur} \quad \partial S_1 \cap \Gamma, \\ v_i &= z_{i-1}^h \quad \text{sur} \quad \partial S_1 \setminus \Gamma \end{aligned} \quad (6.58)$$

par la méthode (6.42) à (6.47). On calcule la solution approchée v_i^h aux nœuds de $\partial D_h \cap \Omega_h$ seuls. L'affirmation du lemme 6.2 implique l'inégalité

$$\max_{\partial D_h \cap \Omega_h} |v_i^h - u| \leqslant d_7 h^2 + q \max_{0 \leqslant j \leqslant N} |z_{i-1}^h(a, \varphi_j) - u(a, \varphi_j)|. \quad (6.59)$$

Le nombre d'opérations arithmétiques est à ce pas de l'ordre de N^3 ou N^2 selon que le second membre est non nul ou $\equiv 0$. Il y a donc intérêt à calculer celui-ci avant la première itération. Certaines variantes de la transformation de Fourier rapide permettent de réduire ce nombre à $N^2 \ln N$ (voir [112]).

ETAPE II. Connaissant les valeurs de v_i^h sur $\partial D_h \cap \Omega_h$, on pose le problème aux différences

$$\begin{aligned} -L_h u_i^h &= f \quad \text{sur} \quad D_h, \\ u_i^h &= g \quad \text{sur} \quad \partial D_h \cap \Gamma_h, \\ u_i^h &= v_i^h \quad \text{sur} \quad \partial D_h \cap \Omega_h. \end{aligned} \quad (6.60)$$

Sa solution vérifie par le lemme 6.1 l'estimation

$$\max_{\overline{D}_h} |u_i^h - u| \leqslant c_3 h^2 + \max_{\partial D_h \cap \Omega_h} |v_i^h - u|. \quad (6.61)$$

ETAPE III. Le procédé continue si l'on connaît les valeurs approchées de u aux nœuds (a, φ_j) du réseau polaire. On effectue donc une interpolation linéaire par morceaux simple sur ces nœuds

à partir des valeurs de u_i^h en trois nœuds les plus proches du réseau carré $\overline{D}_h$, il vient la fonction

$$z_i^h(a, \varphi_j) = \sum_{x \in \overline{D}_h} G_j(x)\, u_i^h(x), \qquad j = 0, 1, \ldots, N,$$

avec les poids $G_j(x)$ qui sont pour tout j positifs en ces trois nœuds de $\overline{D}_h$ et nuls partout ailleurs. Comme la distance de la droite $r = a$ et du sommet de l'angle est une quantité positive, la fonction u est suffisamment régulière dans le voisinage de $r = a$, et on estime

$$\left| u(a, \varphi_j) - \sum_{x \in \overline{D}_h} G_j(x)\, u(x) \right| \leqslant d_8 h^2.$$

Si l'on tient compte de la propriété

$$\left| \sum_{x \in \overline{D}_h} G_j(x) \right| = 1$$

des coefficients d'interpolation, on aboutit aux inégalités

$$|z_i^h(a, \varphi_j) - u(a, \varphi_j)| \leqslant$$
$$\leqslant \sum_{x \in \overline{D}_h} G_j(x)\, |u_i^h(x) - u(x)| +$$
$$+ \left| u(a, \varphi_j) - \sum_{x \in \overline{D}_h} G_j(x)\, u(x) \right|,$$

d'où

$$\max_{0 \leqslant j \leqslant N} |z_i^h(a, \varphi_j) - u(a, \varphi_j)| \leqslant d_8 h^2 + \max_{x \in \overline{D}_h} |u_i^h - u|. \tag{6.62}$$

Ainsi, on vient de décrire le i-ième pas itératif. On vérifie aisément que les inégalités (6.59), (6.61), (6.62) conduisent à

$$\max_{0 \leqslant j \leqslant N} |z_i^h(a, \varphi_j) - u(a, \varphi_j)| \leqslant (c_8 + d_7 + d_8) h^2 +$$
$$+ q \max_{0 \leqslant j \leqslant N} |z_{i-1}^h(a, \varphi_j) - u(a, \varphi_j)|. \tag{6.63}$$

Il suffit de poser au départ $z_0^h \equiv 0$. On recommence plusieurs fois les trois étapes et on aboutit à une solution approchée exacte à l'ordre 2.

THÉORÈME 6.3. *Si le problème* (6.1), (6.2) *remplit les conditions* (6.3) *à* (6.5), *le procédé par itérations* I *à* III *converge, et l'on a pour l'approximation initiale* $z_0^h \equiv 0$

$$\max_{\overline{D}_h} |u_i^h - u| \leqslant d_9 h^2 + q^i \max_{\overline{\Omega}} |u|, \tag{6.64}$$

avec q défini par la relation (6.54).

DÉMONSTRATION. Comme $z_0^h \equiv 0$, l'inégalité (6.63) implique

$$\max_{0 \leqslant j \leqslant N} | z_1^h (a, \varphi_j) | - u (a, \varphi_j) | \leqslant (c_3 + d_7 + d_8) h^2 + q \max_{\bar{\Omega}} | u |. \quad (6.65)$$

Avec $i - 2$ autres estimations (6.63), on a

$$\max_{0 \leqslant j \leqslant N} | z_{i-1}^h (a, \varphi_j) - u (a, \varphi_j) | \leqslant$$

$$\leqslant (d_7 + d_8 + c_3) h^2 \sum_{i=0}^{i-2} q^i + q^{i-1} \max_{\bar{\Omega}} | u | \leqslant$$

$$\leqslant \frac{c_3 + d_7 + d_8}{1 - q} h^2 + q^{i-1} \max_{\bar{\Omega}} | u |. \quad (6.66)$$

Etant donné que la solution approchée du problème (6.58) admet l'estimation (6.59) et celle de (6.60) vérifie (6.61), il découle de (6.66)

$$\max_{\bar{D}_h} | u_i^h - u | \leqslant \left(c_3 + d_7 + \frac{q (c_3 + d_7 + d_8)}{1 - q} \right) h^2 + q^i \max_{\bar{\Omega}} | u |.$$

On pose

$$d_9 = c_3 + d_7 + \frac{q (c_3 + d_7 + d_8)}{1 - q},$$

ce qui donne le résultat voulu (6.64).

Ainsi, les termes de (6.64) deviennent de même ordre au bout de l itérations

$$l = 2 \left[\frac{\ln h}{\ln q} \right]. \quad (6.67)$$

La précision obtenue sur la solution numérique associée aux nœuds de $\bar{D}_h$ est finalement de l'ordre de h^2.

Voici un exemple numérique. On suppose que Ω est l'intérieur d'un carré de côté $b = 2{,}4$, centré en l'origine des coordonnées et muni d'une coupure d'extrémités (0, 0) et $(b/2, 0)$ (fig. 4.10). On demande dans Ω la solution de l'équation de Laplace

$$-\Delta u = 0, \quad (6.68)$$

avec la condition aux limites

$$u = 0 \quad (6.69)$$

sur la coupure Γ_1 et la condition

$$u (x, y) = g (x, y) \quad (6.70)$$

sur les côtés Γ_2 du carré. Ici

$$g(x, y) = \sqrt{\sqrt{x^2 + y^2} - x}.$$

On établit facilement que le problème (6.68) à (6.70) admet pour solution la fonction

$$u(x, y) = g(x, y) \quad \text{sur} \quad \bar{\Omega}.$$

Le problème a été abordé de deux manières différentes. L'un des procédés a consisté à employer un schéma usuel à cinq points de la forme (6.7), et l'autre a suivi les raisonnements de ce paragraphe et tenu compte de la singularité en (0.0). On note que le domaine Ω possède six autres angles. Comme ils sont droits, les conditions de concordance (3.6) sont vérifiées en ces points. Or, ces conditions entraînent, par suite de [8], [51], la propriété pour les dérivées quatrièmes de la solution u du problème (6.68) à (6.70), d'être bornées au voisinage desdits angles. Aussi il est inutile d'appliquer à ceux-ci l'algorithme avec dégagement des singularités qu'on vient de construire. On a regroupé dans le tableau 4.1 les erreurs maxima sur les solutions associées à Ω_h obtenues par les deux méthodes.

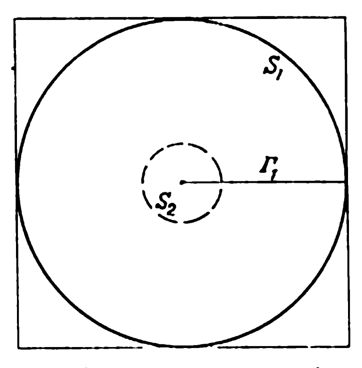


Fig. 4.10. Carré muni d'une coupure

Tableau 4.1

h	0,2	0,15	0,1	0,075	0,05	0,0375
Procédé itératif I à III	0,01444	0,00703	0,00367	0,00206	0,00091	0,00052
Schéma sux différences à cinq points	0,10787	0,09636	0,08158	0,07168	0,05905	0,05147

On note que dans le premier cas l'erreur diminue proportionnellement à h^2, comme il fallait s'y attendre, et que dans le second, elle est à décroissance lente.

CHAPITRE 5

PROBLÈMES NON STATIONNAIRES

Les équations non stationnaires, cas plus ardu que les équations stationnaires, trouvent de nombreuses applications dans diverses branches scientifiques et techniques. La construction d'algorithmes correspondants procède des variables spatiales et aussi du temps, ce qui complique singulièrement le problème d'améliorer la solution approchée obtenue pour plusieurs réseaux successifs. Mais ces améliorations sont néanmoins possibles dans de nombreux cas. C'est justement cette question qui sera traitée dans ce chapitre. Les auteurs passent outre aux schémas de haute précision dont l'idée de base et la mise en œuvre numérique sont différentes, mais qui n'en fournissent pas moins de bons résultats (voir [2], [4], [17], [41], [43], [50], [112], [141]).

5.1. Equation parabolique simple

On se place dans le cas parabolique simple qu'est l'équation de la chaleur en dimension un :

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(p\,\frac{\partial u}{\partial x}\right) + f, \qquad p(x, t) \geq c_0 > 0, \tag{1.1}$$

où $x \in (0, 1)$, $t \in (0, T)$. On définit pour u les conditions initiales et les conditions aux limites

$$u(x, 0) = 0, \qquad x \in [0, 1], \tag{1.2}$$

$$\left.\begin{aligned} u(0, t) &= u_0(t), \\ u(1, t) &= u_1(t), \end{aligned}\right\} \quad t \in [0, T). \tag{1.3}$$

Soit Q le rectangle $(0, 1) \times (0, T)$, $\overline{Q}$ sa fermeture et l un nombre non entier positif quelconque. On introduit une classe de fonctions (voir [25]) pour caractériser la régularité des données initiales, des conditions aux limites et de la solution du problème (1.1) à (1.3).

On appelle $H^l(\bar{Q})$ l'espace de Banach des fonctions $u(x, t)$ qui sont continues sur $\bar{Q}$, ainsi que leurs dérivées de la forme

$$\frac{\partial^{r+s} u}{\partial t^r \partial x^s}, \qquad 2r + s \leq l,$$

pour la norme finie

$$\| u \|_{H^l(\bar{Q})} = \langle u \rangle_Q^{(l)} + \sum_{j=0}^{[l]} \langle u \rangle_Q^{(j)}, \tag{1.4}$$

où

$$\langle u \rangle_Q^{(j)} = \sum_{2r+s=j} \max_{\bar{Q}} \left| \frac{\partial^{r+s} u}{\partial t^r \partial x^s} \right|, \qquad j = 0, 1, \ldots, [l],$$

$$\langle u \rangle_Q^{(l)} = \langle u \rangle_x^{(l)} + \langle u \rangle_t^{(l/2)},$$

$$\langle u \rangle_x^{(l)} = \sum_{2r+s=[l]} \left\langle \frac{\partial^{r+s} u}{\partial t^r \partial x^s} \right\rangle_x^{l-[l]},$$

$$\langle u \rangle_t^{(l/2)} = \sum_{0<l-2r-s<2} \left\langle \frac{\partial^{r+s} u}{\partial t^r \partial x^s} \right\rangle_t^{\frac{l-2r-s}{2}},$$

$$\langle u \rangle_x^{\alpha} = \sup_{(x,t),\,(x',t) \in \bar{Q}} \frac{| u(x, t) - u(x', t) |}{| x - x' |^{\alpha}}, \qquad \alpha \in (0, 1),$$

$$\langle u \rangle_t^{\beta} = \sup_{(x,t),\,(x,t') \in \bar{Q}} \frac{| u(x, t) - u(x, t') |}{| t - t' |^{\beta}}, \qquad \beta \in (0, 1).$$

Pour que la solution soit continue jusqu'à la frontière, il faut que

$$u_0(0) = u_1(0) = 0. \tag{1.5}$$

Nous dirons des égalités (1.5) que ce sont les *conditions de concordance d'ordre* 0. Si l'on veut que la solution admet des dérivées $\partial u/\partial t$ et $\partial^2 u/\partial x^2$ continues jusqu'à la frontière, on exige que

$$\frac{du_0}{dt}(0) = f(0, 0), \qquad \frac{du_1}{dt}(0) = f(1, 0) \tag{1.6}$$

(*conditions de concordance du premier ordre*). On dérive l'équation sous l'hypothèse de continuité jusqu'à la frontière de chaque dérivée, il vient pour $t = 0$:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial}{\partial x} p \frac{\partial u}{\partial x} + f\right) = \frac{\partial}{\partial x}\left(\frac{\partial p}{\partial t} \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial x} p \frac{\partial}{\partial x} \frac{\partial u}{\partial t} + \frac{\partial f}{\partial t} =$$

$$= \frac{\partial}{\partial x} p \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} p \frac{\partial u}{\partial x} + f\right) + \frac{\partial f}{\partial t} - \frac{\partial}{\partial x} p \frac{\partial f}{\partial x} + \frac{\partial f}{\partial t}. \tag{1.7}$$

La fonction u a ses dérivées par rapport à x nulles parce que la condition initiale est homogène. Comme $\partial^2 u/\partial t^2$ aux points anguleux (0, 0) et (1, 0) est exprimée par des quantités connues, la relation (1.7) entraîne

$$\begin{aligned} \frac{d^2 u_0}{dt^2}(0) &= \frac{\partial}{\partial x} p \frac{\partial f}{\partial x}(0, 0) + \frac{\partial f}{\partial t}(0, 0), \\ \frac{d^2 u_1}{dt^2}(0) &= \frac{\partial}{\partial x} p \frac{\partial f}{\partial x}(1, 0) + \frac{\partial f}{\partial t}(1, 0) \end{aligned} \tag{1.8}$$

(*condition de concordance d'ordre* 2). On procède de même pour les conditions d'ordre supérieur. Il se trouve (voir [25]) qu'outre que ces conditions sont nécessaires pour la continuité jusqu'à la frontière des dérivées correspondantes, elles sont suffisantes à condition que les données du problème soient régulières. On énonce cette affirmation sous forme de

THÉORÈME 1.1 [25]. *Soit $k > 0$ non entier,*

$$f \in H^k(\bar{Q}), \qquad p \in H^{k+1}(\bar{Q}), \qquad u_0, u_1 \in C^{k/2+1}[0, T].$$

Si l'on est dans les conditions de concordance d'ordre 0, 1, ..., $[k/2] + 1$, *le problème* (1.1) *à* (1.3) *possède une solution unique* $u \in H^{k+2}(\bar{Q})$.

Soit maintenant le problème (1.1) à (1.3), où $u_0 = u_1 = 0$ sur $[0, T]$, sous les hypothèses du théorème 1.1, la constante l étant dans (3, 4). La condition de valeurs frontières et de valeurs initiales nulles n'est pas embarrassante. Si la fonction u était non nulle à l'instant initial: $u(x, 0) = v(x)$, $x \in [0, 1]$, on aboutirait à la condition initiale homogène en remplaçant u inconnue par $w(x, t) = u(x, t) - v(x)$. Si l'égalité (1.2) est vérifiée, on rend homogènes les conditions (1.3) en posant

$$w(x, t) = u(x, t) - u_0(t)(1 - x) - x u_1(t).$$

Avec w ainsi construite, les conditions (1.2), (1.3) deviennent homogènes.

On fixe $M \geqslant 2$ entier et on pose $\tau = 1/M$. On introduit les notations

$$\bar{\omega}_\tau = \{t_j = j\tau;\ j = 0, 1, \ldots, M\},$$
$$\omega_\tau = \{t_j = j\tau;\ j = 1, 2, \ldots, M\}.$$

On prend le pas d'espace régulier. Soit $h = 1/N$, auquel cas

$$\omega_h = \{x_i = ih;\ i = 1, 2, \ldots, N - 1\},$$
$$\bar{\omega}_h = \{x_i = ih;\ i = 0, 1, \ldots, N\}.$$

On construit les réseaux rectangulaires bidimensionnels en tant que produit cartésien des réseaux en dimension un :

$$\bar{Q}_h^\tau = \bar{\omega}_h \times \bar{\omega}_\tau, \qquad Q_h^\tau = \omega_h \times \omega_\tau . \tag{1.9}$$

Les solutions approchées seront associées aux nœuds de $\bar{Q}_h^\tau$ et les équations aux différences à ceux de Q_h^τ par le schéma implicite (voir [43], [112], [141])

$$u_{\bar{t}}^{\tau} = (p u_{\bar{x}}^{\tau})_x + f \quad \text{sur} \quad Q_h^\tau. \tag{1.10}$$

Comme il y a moins d'équations que d'inconnues, on fait recours aux conditions initiales et aux limites*

$$u^\tau(x, 0) = 0, \qquad x \in \omega_h . \tag{1.11}$$

$$u^\tau(0, t) = u^\tau(1, t) = 0, \qquad t \in \omega_\tau . \tag{1.12}$$

On trouve dans [43] l'estimation suivante qui caractérise la stabilité de la solution discrète :

$$\max_{x \in \bar{\omega}_h} |u^\tau(x, t_{j+1})| \leq \max_{x \in \bar{\omega}_h} |u^\tau(x, 0)| + \sum_{j'=0}^{j} \tau \max_{x \in \omega_h} |f(x, t_{j'})| .$$

Nous en utiliserons la conséquence simple

$$\max_{\bar{Q}_h^\tau} |u^\tau| \leq T \max_{Q_h^\tau} |f| . \tag{1.13}$$

On rappelle qu'on trouve la solution approchée u^τ par la méthode de balayage qui consiste à passer successivement d'un niveau de temps au niveau suivant. Le problème (1.10) à (1.12) est linéaire en u^τ, si bien que l'unicité est fonction du nombre de solutions du problème homogène. Or, le problème homogène (avec le second membre f nul) n'admet par suite de (1.13) que la solution banale, d'où l'unicité de u^τ.

THÉORÈME 1.2. *On suppose qu'on est, pour le problème* (1.1) *à* (1.3), *dans les conditions du théorème* 1.1, *la constante l étant dans* (3, 4). *La solution du problème discret* (1.10) *à* (1.12) *admet le développement*

$$u^\tau(x, t) = u(x, t) + h^2 v(x, t) + \tau w(x, t) + \\ + (h^l + \tau^{l/2})\, \eta^\tau(x, t), \qquad (x, t) \in \bar{Q}_h^\tau. \tag{1.14}$$

* Le souci de simplicité nous fait omettre l'indice h et écrire u^τ tout court.

avec les fonctions v et w continues sur $\bar{Q}$ et indépendantes de τ et h et la fonction discrète η^τ uniformément bornée:

$$|\eta^\tau(x, t)| \leqslant c_1 \qquad \forall (x, t) \in \bar{Q}_h^\tau . \tag{1.15}$$

DÉMONSTRATION. On prend v pour solution du problème

$$\begin{gathered}\frac{\partial v}{\partial t} = \frac{\partial}{\partial x} p \frac{\partial v}{\partial x} + g_1 \quad \text{sur} \quad Q, \\ v(1, t) = v(0, t) = 0, \qquad t \in [0, T], \\ v(x, 0) = 0, \qquad x \in [0, 1],\end{gathered} \tag{1.16}$$

où

$$g_1 = \frac{1}{24}\left(\frac{\partial^3}{\partial x^3} p \frac{\partial u}{\partial x} + \frac{\partial}{\partial x} p \frac{\partial^3 u}{\partial x^3}\right).$$

On note que les données du problème (1.16) remplissent les conditions du théorème 1.1, avec $k = l - 2$. En particulier, la fonction g_1 est continue et devient nulle au voisinage des points anguleux (0, 0) et (1, 0). On est donc dans les conditions de concordance d'ordre 1. La condition d'ordre 0 est vérifiée automatiquement du moment que les conditions aux limites et initiales sont homogènes. Aussi $v \in H^l(\bar{Q})$.

On prend w en tant que solution du problème

$$\begin{gathered}\frac{\partial w}{\partial t} = \frac{\partial}{\partial x} p \frac{\partial w}{\partial x} + g_2 \quad \text{sur} \quad Q, \\ w(x, 0) = 0, \qquad x \in [0, 1], \\ w(1, t) = v(0, t) = 0, \qquad t \in [0, T],\end{gathered} \tag{1.17}$$

où $g_2 = -\dfrac{1}{2}\dfrac{\partial^2 u}{\partial t^2}$. Comme $g_2 \in H^{l-2}(\bar{Q})$ et on est dans les conditions de concordance d'ordre 0 et 1, on a $w \in H^l(\bar{Q})$ en raison du théorème 1.1.

Les fonctions u^τ, u, v, w sont définies de façon unique aux nœuds du réseau Q_h, si bien qu'on accepte l'égalité (1.14) en qualité de la définition de la fonction discrète η^τ. Il reste à prouver la majoration (1.15). On substitue à u^τ de (1.10) le développement (1.14):

$$\begin{aligned} u_{\bar{t}} + h^2 v_{\bar{t}} + \tau w_{\bar{t}} + (h^l + \tau^{l/2})\, \eta_{\bar{t}}^\tau = \\ = (p u_{\bar{x}})_{\hat{x}} + h^2 (p v_{\bar{x}})_{\hat{x}} + \tau (p w_{\bar{x}})_{\hat{x}} + \\ + (h^l + \tau^{l/2})(p \eta_{\bar{x}}^\tau)_{\hat{x}} + f \quad \text{sur} \quad Q_h^\tau. \end{aligned} \tag{1.18}$$

On utilise les lemmes 1.1 et 1.2, § 7.1, il vient

$$u_{\bar{t}} = \frac{\partial u}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2} + \tau^{l/2}\xi_1,$$

$$v_{\bar{t}} = \frac{\partial v}{\partial t} + \tau^{l/2-1}\xi_2,$$

$$w_{\bar{t}} = \frac{\partial w}{\partial t} + \tau^{l/2-1}\xi_3,$$

$$(pu_{\bar{x}})_{\hat{x}} = \frac{\partial}{\partial x} p \frac{\partial u}{\partial x} + h^2\left(\frac{1}{24}\frac{\partial^3}{\partial x^3} p \frac{\partial u}{\partial x} + \frac{1}{24}\frac{\partial}{\partial x} p \frac{\partial^3 u}{\partial x^3}\right) + h^l\xi_4,$$

$$(pv_{\bar{x}})_{\hat{x}} = \frac{\partial}{\partial x} p \frac{\partial v}{\partial x} + h^{l-2}\xi_5,$$

$$(pw_{\bar{x}})_{\hat{x}} = \frac{\partial}{\partial x} p \frac{\partial w}{\partial x} + h^{l-2}\xi_6,$$

où

$$|\xi_i| \leqslant c_i, \qquad i = 1, \ldots, 6,$$

les constantes c_i étant indépendantes des nœuds de Q_h^τ, de h et τ. On remplace les termes de (1.18) par leurs expressions ci-dessus:

$$\frac{\partial u}{\partial t} + h^2\frac{\partial v}{\partial t} + \tau\left(\frac{1}{2}\frac{\partial^2 u}{\partial t^2} + \frac{\partial w}{\partial t}\right) + (h^2\tau^{l/2-1} + \tau^{l/2})\xi_7 +$$
$$+ (h^l + \tau^{l/2})\,\eta_{\bar{t}}^\tau = \frac{\partial}{\partial x} p \frac{\partial u}{\partial x} + h^2\left(\frac{1}{24}\frac{\partial^3}{\partial x^3} p \frac{\partial u}{\partial x} + \frac{1}{24}\frac{\partial}{\partial x} p \frac{\partial^3 u}{\partial x^3} + \frac{\partial}{\partial x} p \frac{\partial v}{\partial x}\right) +$$
$$+ \tau\frac{\partial}{\partial x} p \frac{\partial w}{\partial x} + (h^l + h^{l-2}\tau)\,\xi_8 + (h^l + \tau^{l/2})(p\eta_{\bar{x}}^\tau)_{\hat{x}} + f \quad \text{sur} \quad Q_h^\tau. \quad (1.19)$$

Ici

$$|\xi_7| \leqslant c_1 + c_2 + c_3, \qquad |\xi_8| \leqslant c_4 + c_5 + c_6.$$

On omet les termes en h^0, τ^0 car u vérifie l'équation (1.1). Les coefficients de h^2 des deux membres de (1.19) se détruisent mutuellement parce que l'équation du problème (1.16) entraîne

$$\frac{\partial v}{\partial t} = \frac{\partial}{\partial x} p \frac{\partial v}{\partial x} + \frac{1}{24}\frac{\partial^3}{\partial x^3} p \frac{\partial u}{\partial x} + \frac{1}{24}\frac{\partial}{\partial x} p \frac{\partial^3 u}{\partial x^3}.$$

La fonction w est telle que

$$\frac{\partial w}{\partial t} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} p \frac{\partial w}{\partial x},$$

condition qui permet d'identifier les termes en τ. La formule (1.19) se récrit donc

$$(h^l + \tau^{l/2})\,\eta_{\bar{t}}^\tau = (h^l + \tau^{l/2})\,(p\eta_{\bar{x}}^\tau)_{\hat{x}} + (h^l + h^{l-2}\tau)\,\xi_8 +$$
$$+ (h^2\tau^{l/2-1} + \tau^{l/2})\,\xi_7 \quad \text{sur} \quad Q_h^\tau. \quad (1.20)$$

On fait appel à l'inégalité de Young (voir [26]) : si $a \geq 0$, $b \geq 0$, alors

$$ab \leqslant \frac{1}{p} a^p + \frac{1}{p'} b^{p'},$$

où

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

S'agissant des produits $h^{l-2}\tau$ et $h^2 \tau^{l/2-1}$, on a

$$h^{l-2} \tau \leqslant \left(1 - \frac{2}{l}\right) h^l + \frac{2}{l} \tau^{l/2} \leqslant h^l + \tau^{l/2},$$

$$h^2 \tau^{l/2-1} \leqslant \frac{2}{l} h^l + \left(1 - \frac{2}{l}\right) \tau^{l/2} \leqslant h^l + \tau^{l/2}$$

respectivement. On divise (1.20) membre à membre par $h^l + \tau^{l/2}$ et on utilise deux dernières inégalités, il vient

$$\eta_t^\tau = (p\eta_{\bar{x}}^\tau)_x + \xi_9 \quad \text{sur} \quad Q_h^\tau, \tag{1.21}$$

avec

$$|\xi_9| \leqslant 2 \sum_{i=1}^{6} c_i = c_7. \tag{1.22}$$

Afin d'appliquer l'estimation (1.13), il faut déterminer les conditions initiales et aux limites pour η^τ discrète. On a pour $t = 0$

$$u(x, 0) = v(x, 0) = w(x, 0) = 0, \qquad x \in [0, 1],$$
$$u^\tau(x, 0) = 0, \qquad x \in \bar{\omega}_h,$$

si bien que la formule (1.14) implique

$$\eta^\tau(x, 0) = 0, \qquad x \in \bar{\omega}_h.$$

On vérifie de même à l'aide des conditions aux limites correspondantes que

$$\eta^\tau(0, t) = \eta^\tau(1, t) = 0 \qquad \forall t \in \omega_\tau.$$

Ainsi, l'estimation (1.13) est valable pour l'équation (1.21) et fournit l'inégalité

$$\max_{\bar{Q}_h^\tau} |\eta^\tau| \leqslant T \max_{Q_h^\tau} |\xi_9| \leqslant c_7 T.$$

On a donc la majoration (1.15), où $c_1 = c_7 T$, ce qui achève la démonstration du théorème.

On note que la régularité en x et en t de la solution n'est pas la même, si bien que les pas h et τ ne sont pas équivalents en ce qui concerne l'ordre de précision. Si la régularité en x entraîne que la précision de la solution approchée est au maximum en h^k, elle est

en $\tau^{k/2}$ lorsqu'on fait abstraction des dérivées d'espace. Il est donc naturel que le développement (1.14) renferme τ à côté de h^2.

Nous voulons décrire un procédé de raffinement basé sur le développement (1.14). On admet que toutes les conditions du théorème 1.2 sont remplies. On choisit $M \geqslant 2$ et $N \geqslant 2$ entiers et on construit le réseau $\overline{Q}_h^\tau$ de pas temporel $\tau = 1/M$ et de pas d'espace $h = 1/N$. On cherche la solution u^τ du problème (1.10) à (1.12) sur le réseau $\overline{Q}_h^\tau$.

On construit le réseau $\overline{Q}_{h/2}^{\tau/4}$ de pas $\tau/4$ et $h/2$ et on cherche la solution $u^{\tau/4}$ du problème. Ainsi, on a deux solutions approchées u^τ et $u^{\tau/4}$ associées aux nœuds de $\overline{Q}_h^\tau$. On prend leur combinaison linéaire

$$U(x, t) = \frac{4}{3}\, u^{\tau/4}(x, t) - \frac{1}{3}\, u^\tau (x, t), \qquad (x, t) \in \overline{Q}_h^\tau \quad (1.23)$$

et on montre que la solution améliorée U approche la solution exacte avec une précision en $h^l + \tau^{l/2}$ pour la métrique uniforme.

Les solutions approchées admettent en chaque nœud $(x, t) \in \overline{Q}_h^\tau$ les représentations

$$u^\tau (x, t) = u(x, t) + h^2 v(x, t) + \tau w(x, t) + (h^l + \tau^{l/2})\, \eta_l^\tau (x, t),$$

$$u^{\tau/4}(x, t) = u(x, t) + \frac{h^2}{4} v(x, t) + \frac{\tau}{4} w(x, t) + \frac{h^l + \tau^{l/2}}{2^l}\, \eta_l^{\tau/4} (x, t).$$

Comme les fonctions u, v et w sont indépendantes de τ et h, on utilise l'égalité de leurs valeurs qui fait que les termes en v et w se détruisent mutuellement, et on forme la combinaison linéaire:

$$U(x, t) = u(x, t) + (h^l + \tau^{l/2})\, \frac{1}{3}\, (2^{2-l} \eta_l^{\tau/4} (x, t) - \eta_l^\tau (x, t)).$$

Etant donnée l'estimation (1.15), on trouve

$$|U(x, t) - u(x, t)| \leqslant$$

$$\leqslant \frac{1}{3} (2^{2-l} + 1)\, c_1 (h^l + \tau^{l/2}) \leqslant \frac{c_1}{2} (h^l + \tau^{l/2}) \quad (1.24)$$

pour tout $(x, t) \in \overline{Q}_h^\tau$.

Ainsi, la solution (1.23), combinaison linéaire des solutions approchées exactes à l'ordre $\tau + h^2$ approche la solution correcte $u(x, t)$ aux nœuds de $\overline{Q}_h^\tau$ avec une précision en $h^l + \tau^{l/2}$, $3 < l < 4$.

Il y a lieu de dire que quitte d'augmenter légèrement la régularité des fonctions f et p (par exemple, $f \in H^{4+\alpha} (\overline{Q})$, $p \in H^{5+\alpha} (\overline{Q})$, où $\alpha \in (0, 1)$, tout en se contentant des conditions de concordance du théorème 1.1, on garantit pour (1.23) une précision de l'ordre de $h^4 + \tau^2$. En effet, les dérivées figurant dans les restes en h^4 et τ^2 des erreurs d'approximation sont bornées malgré les discontinuités

aux points anguleux (0, 0) et (1, 0). On le justifie en général par de longs calculs qui n'ont aucun rapport avec le schéma aux différences concerné.

Si l'on exige que les données du problème soient plus régulières et que le nombre de conditions de concordance dépasse celui du théorème 1.1, alors la solution du problème (1.1) à (1.3) présente une régularité plus grande. Il se trouve cependant qu'on n'obtient un développement (1.14) suivant les puissances de h^2 et τ avec reste d'ordre supérieur à $h^2 + \tau$ qu'en modifiant l'équation différentielle. Le fait est que les coefficients du développement sont définis à partir des problèmes auxiliaires qui ne satisfont qu'à deux conditions de concordance. Pour qu'elles soient plus nombreuses, on fait la substitution

$$u(x, t) = z(x, t) + tx(1 - x) R(x, t).$$

Ici z est la nouvelle fonction inconnue et R un polynôme en x et t qui garantit la concordance pour les problèmes auxiliaires. On note que cette substitution n'influe pas sur la régularité de la solution ni sur l'homogénéité des conditions initiales et aux limites.

Si l'on approche les dérivées d'espace par des schémas aux différences du chapitre 3, on adapte assez facilement les résultats obtenus pour les coefficients discontinus (le lieu de discontinuité étant fixé), le troisième problème aux limites et les coefficients dépendant de la solution. (Dans [43] l'estimation utile (1.13) est justement transposée à ce cas.)

5.2. Amélioration de la précision dans une méthode de décomposition

Soit l'équation d'évolution abstraite (voir [21])

$$\frac{d\mathbf{u}}{dt} + A\mathbf{u} = \mathbf{f}, \qquad t \in (0, T). \tag{2.1}$$

avec la condition initiale

$$\mathbf{u}(0) = \mathbf{u}_0. \tag{2.2}$$

Ici $\mathbf{u}(t)$ et $\mathbf{f}(t)$ sont éléments d'un espace de Hilbert H muni du produit scalaire $(\mathbf{u}, \mathbf{v})$ et de la norme

$$\|\mathbf{u}\| = (\mathbf{u}, \mathbf{u})^{1/2}.$$

L'opérateur linéaire A est indépendant du temps* et défini non négatif, i.e.

$$(A\mathbf{u}, \mathbf{u}) \geq 0 \qquad \forall \mathbf{u} \in H.$$

* L'indépendance par rapport à t, voire à $\mathbf{u}$, est sans importance.

On suppose de plus que la solution $\mathbf{u}$ du problème et le second membre $\mathbf{f}$ possèdent un nombre suffisant de dérivées bornées par rapport à t. Nous nous proposons de justifier de deux manières différentes la convergence et l'augmentation de la précision pour une méthode de décomposition.

L'opérateur A est supposé admettre la représentation $A = A_1 + A_2$, A_1 et A_2 étant non négatifs. On partage le segment $[0, T]$ en M parties égales et on introduit les notations

$$\bar{\omega}_\tau = \{t_j = j\tau;\ j = 0, 1, \ldots, M\},$$

$$\omega_\tau = \{t \in \bar{\omega}_\tau,\ t \neq 0\}.$$

Soit le schéma

$$\begin{aligned} &\frac{\mathbf{u}^*(t) - \mathbf{u}^\tau(t-\tau)}{\tau} + A_1 \mathbf{u}^*(t) = \mathbf{f}(t), \\ &\frac{\mathbf{u}^\tau(t) - \mathbf{u}^*(t)}{\tau} + A_2 \mathbf{u}^\tau(t) = 0, \quad t \in \omega_\tau, \end{aligned} \tag{2.3}$$

avec la condition initiale

$$\mathbf{u}^\tau(0) = \mathbf{u}_0. \tag{2.4}$$

Le schéma est absolument stable, et on a l'estimation à priori

$$\max_{t \in \bar{\omega}_\tau} \| \mathbf{u}^\tau(t) \| \leqslant \| \mathbf{u}_0 \| + T \max_{t \in \omega_\tau} \| \mathbf{f}(t) \| \tag{2.5}$$

qui garantit l'unicité et la propriété de borne de la solution approchée (voir [30]).

La première façon d'agir pour légitimer le procédé de raffinement consiste à éliminer la valeur intermédiaire et à examiner les propriétés d'approximation et la convergence pour le problème réduit

$$\begin{aligned} (I + \tau A_1)(I + \tau A_2)\, \mathbf{u}^\tau(t) &= \mathbf{u}^\tau(t - \tau) + \tau \mathbf{f}(t), \quad t \in \omega_\tau, \\ \mathbf{u}^\tau(0) &= \mathbf{u}_0. \end{aligned} \tag{2.6}$$

Ici I est l'opérateur identité dans H. On suppose d'abord que

$$\mathbf{u}^\tau = \mathbf{u} + \tau \mathbf{v}_1 + \tau^2 \mathbf{v}_2 + \tau^3 \boldsymbol{\eta}^\tau \quad \text{sur} \quad \bar{\omega}_\tau, \tag{2.7}$$

avec $\mathbf{v}_1$, $\mathbf{v}_2$ indépendantes de τ et $\boldsymbol{\eta}^\tau$ une fonction discrète bornée. On substitue ce développement à $\mathbf{u}^\tau$ de (2.6), il vient

$$(I + \tau A_1)(I + \tau A_2)(\mathbf{u} + \tau \mathbf{v}_1 + \tau^2 \mathbf{v}_2 + \tau^3 \boldsymbol{\eta}^\tau)|_t =$$

$$= (\mathbf{u} + \tau \mathbf{v}_1 + \tau^2 \mathbf{v}_2 + \tau^3 \boldsymbol{\eta}^\tau)|_{t-\tau} + \tau \mathbf{f}(t).$$

On développe $\mathbf{u}$, $\mathbf{v}_1$, $\mathbf{v}_2$ supposées suffisamment régulières en formule de Taylor

$$\mathbf{u}(t-\tau) = \mathbf{u}(t) - \tau \frac{d\mathbf{u}}{dt}(t) + \frac{\tau^2}{2}\frac{d^2\mathbf{u}}{dt^2}(t) - \frac{\tau^3}{6}\frac{d^3\mathbf{u}}{dt^3}(t) + \tau^4 \xi_1,$$

$$\mathbf{v}_1(t-\tau) = \mathbf{v}_1(t) - \tau\frac{d\mathbf{v}_1}{dt}(t) + \frac{\tau^2}{2}\frac{d^2\mathbf{u}}{dt^2}(t) + \tau^3 \xi_2, \tag{2.8}$$

$$\mathbf{v}_2(t-\tau) = \mathbf{v}_2(t) - \tau\frac{d\mathbf{v}_2}{dt}(t) + \tau^2 \xi_3,$$

où

$$\|\xi_i\| \leqslant c_1 \qquad \forall t \in \omega_\tau, \ i = 1, 2, 3.$$

On a

$$(I+\tau A_1)(I+\tau A_2)(\mathbf{u}+\tau\mathbf{v}_1+\tau^2\mathbf{v}_2+\tau^3\boldsymbol{\eta}^\tau)|_t =$$
$$= \mathbf{u}(t) + \tau\left(-\frac{d\mathbf{u}}{dt}+\mathbf{v}_1\right)\Big|_t + \tau^2\left(\frac{1}{2}\frac{d^2\mathbf{u}}{dt^2}-\frac{d\mathbf{v}_1}{dt}+\mathbf{v}_2\right)\Big|_t +$$
$$+\tau^3\left(-\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3}+\frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2}-\frac{d\mathbf{v}_2}{dt}\right)\Big|_t + \tau^3\boldsymbol{\eta}^\tau(t-\tau) +$$
$$+\tau\mathbf{f}(t) + \tau^3(\xi_1+\xi_2+\xi_3)|_t .$$

La multiplication des parenthèses et la réduction des termes semblables donnent

$$\tau^2(A\mathbf{v}_1 + A_1A_2\mathbf{u})|_t + \tau^3(A\mathbf{v}_2 + A_1A_2\mathbf{v}_1)|_t +$$
$$+\tau^4 A_1A_2\mathbf{v}_2(t) + \tau^3(I+\tau A_1)(I+\tau A_2)\boldsymbol{\eta}^\tau(t) =$$
$$= \tau_2\left(\frac{1}{2}\frac{d^2\mathbf{u}}{dt^2}-\frac{d\mathbf{v}_1}{dt}\right)\Big|_t + \tau^3\left(-\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3}+\frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2}-\frac{d\mathbf{v}_2}{dt}\right)\Big|_t +$$
$$+\tau^3\boldsymbol{\eta}^\tau(t-\tau) + \tau^4(\xi_1+\xi_2+\xi_3)|_t . \tag{2.9}$$

On égale les coefficients de τ^2:

$$A\mathbf{v}_1 + A_1A_2\mathbf{u} = \frac{1}{2}\frac{d^2\mathbf{u}}{dt^2} - \frac{d\mathbf{v}_1}{dt}, \qquad t\in\omega_\tau .$$

On fait l'hypothèse de $A_1A_2\mathbf{u}$ borné pour tout $t\in[0, T]$, auquel cas on cherche $\mathbf{v}_1$ comme solution du problème

$$\frac{d\mathbf{v}_1}{dt} + A\mathbf{v}_1 = \frac{1}{2}\frac{d^2\mathbf{u}}{dt^2} - A_1A_2\mathbf{u}, \qquad t\in[0, T],$$
$$\mathbf{v}_1(0) = 0. \tag{2.10}$$

La condition initiale exige quelques éclaircissements. Le développement (2.7) a lieu sur $\bar{\omega}_\tau$ si l'on a à l'instant initial

$$\mathbf{u}^\tau(0) = \mathbf{u}(0) + \tau\mathbf{v}_1(0) + \tau^2\mathbf{v}_2(0) + \tau^3\boldsymbol{\eta}^\tau(0). \tag{2.11}$$

Puisque $\mathbf{u}^\tau(0) = \mathbf{u}(0) = \mathbf{u}_0$, l'identification des termes semblables donne

$$\tau\, \mathbf{v}_1(0) + \tau^2\, \mathbf{v}_2(0) + \tau^3\, \boldsymbol{\eta}^\tau(0) = \mathbf{0}. \tag{2.12}$$

La condition initiale du problème (2.10) vient en égalant à 0 le coefficient de τ.

Le fait d'avoir pris $\mathbf{v}_1$ pour solution de (2.10) ramène la relation (2.9) à

$$(A\, \mathbf{v}_2 + A_1 A_2 \mathbf{v}_1)|_t + \tau A_1 A_2 \mathbf{v}_2(t) + (I + \tau A_1)(I + \tau A_2)\, \boldsymbol{\eta}^\tau(t) =$$
$$= -\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3}(t) + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2}(t) - \frac{d\mathbf{v}_2}{dt}(t) + \boldsymbol{\eta}^\tau(t-\tau) +$$
$$+ \tau(\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + \boldsymbol{\xi}_3)|_t. \tag{2.13}$$

Pour que la fonction $\boldsymbol{\eta}^\tau$ soit bornée, il suffit, conformément à (2.5), de choisir $\mathbf{v}_2$ telle que (2.13) ne renferme que les termes en τ et les termes contenant $\boldsymbol{\eta}^\tau$. Ce choix de v_2 conduit à l'égalité

$$A\, \mathbf{v}_2 + A_1 A_2\, \mathbf{v}_1 = -\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2} - \frac{d\mathbf{v}_2}{dt}, \quad t \in \omega_\tau.$$

On assujettit $\mathbf{v}_2$ à satisfaire à

$$\frac{d\mathbf{v}_2}{dt} + A\, \mathbf{v}_2 = -\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2} - A_1 A_2 \mathbf{v}_1, \quad t \in [0, T],$$
$$\mathbf{v}_2(0) = \mathbf{0} \tag{2.14}$$

sur tout le segment. On note que la condition initiale homogène du problème de Cauchy ainsi construit garantit la borne de $\boldsymbol{\eta}^\tau(0)$ de (2.12). Avec l'hypothèse de $A_1 A_2 \mathbf{v}_1$ borné, on aboutit donc au problème qui définit $\mathbf{v}_2$ de façon unique.

La relation (2.13) devient pour $\mathbf{v}_1$ et $\mathbf{v}_2$ ainsi choisis

$$(I + \tau A_1)(I + \tau A_2)\, \boldsymbol{\eta}^\tau(t) =$$
$$= \boldsymbol{\eta}^\tau(t-\tau) + \tau(\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + \boldsymbol{\xi}_3)|_t - \tau A_1 A_2 \mathbf{v}_2(t), \quad t \in \omega_\tau. \tag{2.15}$$

Ainsi, les fonctions $\mathbf{u}^\tau$, $\mathbf{u}$, $\mathbf{v}_1$, $\mathbf{v}_2$ sont définies dans (2.7) de façon unique. Il en est donc de même de $\boldsymbol{\eta}^\tau$. On démontre la propriété de cette fonction d'être bornée. L'égalité (2.12) donne lieu à la condition initiale

$$\boldsymbol{\eta}^\tau(0) = \mathbf{0}. \tag{2.16}$$

Le produit $A_1 A_2 \mathbf{v}_2$ étant borné, on a par suite de l'estimation (2.5)

$$\|\boldsymbol{\eta}^\tau\| \leqslant c_2 \qquad \forall\, t \in \bar{\omega}_\tau. \tag{2.17}$$

Ainsi, les fonctions $\mathbf{v}_1$, $\mathbf{v}_2$ sont définies et indépendantes de τ, et la fonction discrète $\boldsymbol{\eta}^\tau$ est bornée.

Comment on utilise le développement obtenu ? Soit $M > 2$ un entier naturel. On cherche les solutions du problème (2.3), (2.4) pour les pas $\tau = T/M$, $\tau/2$, $\tau/3$. On a pour le réseau $\bar{\omega}_\tau$ de pas τ :

$$\mathbf{u}^\tau = \mathbf{u} + \tau\,\mathbf{v}_1 + \tau^2\,\mathbf{v}_2 + \tau^3\,\boldsymbol{\eta}^\tau,$$

$$\mathbf{u}^{\tau/2} = \mathbf{u} + \frac{\tau}{2}\,\mathbf{v}_1 + \frac{\tau^2}{4}\,\mathbf{v}_2 + \frac{\tau^3}{8}\,\boldsymbol{\eta}^{\tau/2},$$

$$\mathbf{u}^{\tau/3} = \mathbf{u} + \frac{\tau}{3}\,\mathbf{v}_1 + \frac{\tau^2}{9}\,\mathbf{v}_2 + \frac{\tau^3}{27}\,\boldsymbol{\eta}^{\tau/3}.$$

On fait la somme avec les poids respectifs $\gamma_1 = 1/2$, $\gamma_2 = -4$, $\gamma_3 = 9/2$, il vient

$$\mathbf{U} = \frac{1}{2}\,\mathbf{u}^\tau - 4\,\mathbf{u}^{\tau/2} + \frac{9}{2}\,\mathbf{u}^{\tau/3} = \mathbf{u} + \tau^3\left(\frac{1}{2}\,\boldsymbol{\eta}^\tau - \frac{1}{2}\,\boldsymbol{\eta}^{\tau/2} + \frac{1}{6}\,\boldsymbol{\eta}^{\tau/3}\right),$$

$$t \in \bar{\omega}_\tau.$$

Comme γ_i sont obtenus comme solution du système

$$\gamma_1 + \gamma_2 + \gamma_3 = 1,$$

$$\gamma_1 + \frac{1}{2}\gamma_2 + \frac{1}{3}\gamma_3 = 0,$$

$$\gamma_1 + \frac{1}{4}\gamma_2 + \frac{1}{9}\gamma_3 = 0,$$

les termes en τ et τ^2 dans l'égalité précédente s'annulent. L'estimation (2.17) des restes fournit

$$\|\,\mathbf{U}(t) - \mathbf{u}(t)\,\| \leqslant \tau^3\,\frac{7}{6}\,c_2 \qquad \forall\, t \in \bar{\omega}_\tau. \tag{2.18}$$

Cette inégalité montre le gain en précision apporté par la solution corrigée.

Ce procédé de justification s'avère très rationnel si A est une matrice algébrique ou un opérateur intégral. On l'utilise des fois dans le cas d'opérateurs différentiels. Dans le § 5.4, on l'appliquera à l'équation du mouvement.

La méthode est inopérante pour une classe plus vaste de problèmes parce que le produit $A_1 A_2 \mathbf{w}$ est souvent non borné pour une classe fonctionnelle étendue (nous en avons parlé plus d'une fois). S'agissant des opérateurs différentiels, ce produit n'est en général pas défini pour la plupart des problèmes aux limites. Le fait de discrétiser au préalable le problème par rapport aux variables spatiales ne change rien à la situation. Bien qu'on obtienne une matrice A algébrique et que l'approche décrite soit formellement possible, les grandeurs des pas d'espace interviennent dans toutes les constantes caractéristiques de l'approximation et de la convergence. Cela

altère des fois l'ordre de convergence, et il se peut même que la convergence n'ait pas lieu (voir [141]).

Le produit $A_1 A_2 \mathbf{w}$ est donc à éviter. On procède comme suit. Au lieu d'éliminer la valeur intermédiaire $\mathbf{u}^*(t)$ du schéma (2.3), on l'interprète comme une approximation de la solution. C'est l'idée de base de la *méthode d'approximation additive* (voir [43]), qui conduit à un algorithme de raffinement.

Soit les développements de $\mathbf{u}^\tau$ et $\mathbf{u}^*$

$$\mathbf{u}^\tau = \mathbf{u} + \tau \mathbf{v}_1 + \tau^2 \mathbf{v}_2 + \tau^3 \boldsymbol{\eta}^\tau \quad \text{sur} \quad \bar{\omega}_\tau. \tag{2.19}$$

$$\mathbf{u}^* = \mathbf{u} + \tau \mathbf{w}_1 + \tau^2 \mathbf{w}_2 + \tau^3 \boldsymbol{\zeta}^\tau \quad \text{sur} \quad \omega_\tau. \tag{2.20}$$

Les fonctions $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{w}_1$, $\mathbf{w}_2$ ne dépendent pas de τ, et $\boldsymbol{\eta}^\tau$ et $\boldsymbol{\zeta}^\tau$ sont bornées sur $\bar{\omega}_\tau$.

On se propose de justifier ces développements. On les porte dans les formules (2.3)*:

$$\frac{\mathbf{u} - \mathbf{u}(t-\tau)}{\tau} + \mathbf{w}_1 + \tau \mathbf{w}_2 + \tau^2 \boldsymbol{\zeta}^\tau - \mathbf{v}_1(t-\tau) - \\ - \tau \mathbf{v}_2(t-\tau) - \tau^2 \boldsymbol{\eta}^\tau(t-\tau) + A_1 \mathbf{u} + \tau A_1 \mathbf{w}_1 + \\ + \tau^2 A_1 \mathbf{w}_2 + \tau^3 A_1 \boldsymbol{\zeta}^\tau = \mathbf{f}. \tag{2.21}$$

$$\mathbf{v}_1 + \tau \mathbf{v}_2 + \tau^2 \boldsymbol{\eta}^\tau - \mathbf{w}_1 - \tau \mathbf{w}_2 - \tau^2 \boldsymbol{\zeta}^\tau + \\ + A_2 \mathbf{u} + \tau A_2 \mathbf{v}_1 + \tau^2 A_2 \mathbf{v}_2 + \tau^3 A_2 \boldsymbol{\eta}^\tau = 0. \tag{2.22}$$

Avec les développements tayloriens (2.8) de $\mathbf{u}$, $\mathbf{v}_1$, $\mathbf{v}_2$, la relation (2.21) se récrit

$$\frac{d\mathbf{u}}{dt} - \frac{\tau}{2}\frac{d^2\mathbf{u}}{dt^2} + \frac{\tau^2}{6}\frac{d^3\mathbf{u}}{dt^3} + \mathbf{w}_1 - \mathbf{v}_1 + \tau \frac{d\mathbf{v}_1}{dt} - \frac{\tau^2}{2}\frac{d^2\mathbf{v}_1}{dt^2} + \tau \mathbf{w}_2 - \tau \mathbf{v}_2 + \\ + \tau^2 \frac{d\mathbf{v}_2}{dt} + \tau \boldsymbol{\zeta}^\tau - \tau^2 \boldsymbol{\eta}^\tau(t-\tau) + A_1 \mathbf{u} + \tau A_1 \mathbf{w}_1 + \tau^2 A_1 \mathbf{w}_2 + \\ + \tau^3(\xi_1 + \xi_2 + \xi_3) + \tau^3 A_1 \boldsymbol{\zeta}^\tau = \mathbf{f}, \quad t \in \omega_\tau. \tag{2.23}$$

On identifie les coefficients de τ^0 des formules (2.22), (2.23), il vient

$$\mathbf{v}_1 - \mathbf{w}_1 + A_2 \mathbf{u} = 0, \\ \frac{d\mathbf{u}}{dt} + \mathbf{w}_1 - \mathbf{v}_1 + A_1 \mathbf{u} = \mathbf{f}, \tag{2.24} \\ t \in \omega_\tau.$$

Ce système est dégénéré par rapport aux inconnues $\mathbf{v}_1$, $\mathbf{w}_1$. On élimine la différence $\mathbf{v}_1 - \mathbf{w}_1$ de la deuxième équation à l'aide de la première et on a l'identité

$$\frac{d\mathbf{u}}{dt} + A\mathbf{u} = \mathbf{f}.$$

* Dans la suite la variable indépendante t sera partout omise.

Aussi $\mathbf{v}_1$, $\mathbf{w}_1$ du système (2.24) sont soumis à une seule condition, à savoir

$$\mathbf{w}_1 - \mathbf{v}_1 = A_2\mathbf{u}, \qquad t \in (0, T]. \tag{2.25}$$

On obtient par identification des coefficients de τ

$$\begin{gathered} \mathbf{v}_2 - \mathbf{w}_2 + A_2\mathbf{v}_1 = 0, \\ -\frac{1}{2}\frac{d^2\mathbf{u}}{dt^2} + \frac{d\mathbf{v}_1}{dt} + \mathbf{w}_2 - \mathbf{v}_2 + A_1\mathbf{w}_1 = 0, \\ t \in \omega_\tau. \end{gathered} \tag{2.26}$$

Le système a beau être dégénéré par rapport à $\mathbf{v}_2$ et $\mathbf{w}_2$, il possède une solution sous la condition de compatibilité

$$-\frac{1}{2}\frac{d^2\mathbf{u}}{dt^2} + \frac{d\mathbf{v}_1}{dt} + A_2\mathbf{v}_1 + A_1\mathbf{w}_1 = 0, \qquad t \in \omega_\tau.$$

On exige que cette égalité ait lieu non seulement aux nœuds du réseau, mais aussi sur $(0, T]$ tout entier:

$$\frac{d\mathbf{v}_1}{dt} + A_2\mathbf{v}_1 + A_1\mathbf{w}_1 = \frac{1}{2}\frac{d^2\mathbf{u}}{dt^2}, \qquad t \in (0, T]. \tag{2.27}$$

On note que le système (2.25), (2.27) constitue un problème algébrique différentiel. Pour que le développement (2.19) soit juste en tous les nœuds de $\bar{\omega}_\tau$, il faut avoir à l'instant initial

$$\mathbf{u}^\tau(0) = \mathbf{u}(0) + \tau\mathbf{v}_1(0) + \tau^2\mathbf{v}_2(0) + \tau^3\boldsymbol{\eta}^\tau(0). \tag{2.28}$$

Du moment que $\mathbf{u}^\tau(0) = \mathbf{u}(0) = \mathbf{u}_0$, on a

$$\tau\mathbf{v}_1(0) + \tau^2\mathbf{v}_2(0) + \tau^3\boldsymbol{\eta}^\tau(0) = 0. \tag{2.29}$$

On obtient la condition initiale $\mathbf{v}_1(0) = 0$ en égalant à 0 le coefficient de τ.

On suppose que le problème

$$\begin{gathered} \frac{d\mathbf{v}_1}{dt} + A_2\mathbf{v}_1 + A_1\mathbf{w}_1 = \frac{1}{2}\frac{d^2\mathbf{u}}{dt^2}, \qquad t \in (0, T], \\ \mathbf{w}_1 - \mathbf{v}_1 = A_2\mathbf{u}, \qquad t \in (0, T], \\ \mathbf{v}_1(0) = 0 \end{gathered} \tag{2.30}$$

admet au moins une solution. L'existence et la régularité de la fonction restent pour le moment des problèmes ouverts que nous examinerons dans chaque cas concret. Quant à l'unicité, elle est établie sans difficulté. En effet, l'unicité pour le problème linéaire (2.30) est liée à l'absence de solutions non triviales du problème homogène

$$\begin{gathered} \frac{d\mathbf{v}_1}{dt} + A_2\mathbf{v}_1 + A_1\mathbf{w}_1 = 0 \quad \text{sur} \quad (0, T], \\ \mathbf{w}_1 - \mathbf{v}_1 = 0 \quad \text{sur} \quad (0, T], \\ \mathbf{v}_1(0) = 0. \end{gathered}$$

Or, il ne peut avoir de solution non nulle parce qu'on est ramené par élimination de l'inconnue $\mathbf{w}_1$ à

$$\frac{d\mathbf{v}_1}{dt} + A\mathbf{v}_1 = 0 \quad \text{sur} \quad (0, T],$$
$$\mathbf{v}_1(0) = 0$$

qui ne possède que la solution triviale (voir [21]).

On choisit $\mathbf{v}_1$, $\mathbf{w}_1$ à partir des conditions (2.30), ce qui détruit les termes en τ^2 de (2.22) et (2.23). On divise les termes restants par τ^3, il vient le schéma décomposé usuel à pas locaux implicites

$$\frac{\boldsymbol{\eta}^\tau - \boldsymbol{\zeta}^\tau}{\tau} + A_2 \boldsymbol{\eta}^\tau = -\frac{1}{\tau} A_2 \mathbf{v}_2, \tag{2.31}$$

$$\frac{\boldsymbol{\zeta}^\tau - \boldsymbol{\eta}^\tau(t-\tau)}{\tau} + A_1 \boldsymbol{\zeta}^\tau =$$
$$= \frac{1}{\tau}\left(-\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2} - \frac{d\mathbf{v}_2}{dt} - A_1\mathbf{w}_2\right) + \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + \boldsymbol{\xi}_3,$$
$$t \in \omega_\tau. \tag{2.32}$$

Les conditions

$$A_1 \mathbf{v}_2 = 0,$$
$$-\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2} - \frac{d\mathbf{v}_2}{dt} - A_1\mathbf{w}_2 = 0$$

sont suffisantes pour que les fonctions $\boldsymbol{\eta}^\tau$ et $\boldsymbol{\zeta}^\tau$ soient bornées. Mais le problème de recherche de $\mathbf{v}_2$, $\mathbf{w}_2$ devient surdéterminé (si l'on compte la première équation du système (2.26), il y a trois équations pour deux inconnues) et, ce qui plus est, incompatible. On fait appel à la méthode d'approximation additive. Pour que $\boldsymbol{\eta}^\tau$ et $\boldsymbol{\zeta}^\tau$ soient bornées, il suffit que la somme des seconds membres constitue une fonction bornée. On a

$$-\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2} - \frac{d\mathbf{v}_2}{dt} - A_1\mathbf{w}_2 - A_2\mathbf{v}_2 = 0.$$

On prolonge la relation au segment $(0, T]$ tout entier:

$$\frac{d\mathbf{v}_2}{dt} + A_1\mathbf{w}_2 + A_2\mathbf{v}_2 = -\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2}, \qquad t \in (0, T]. \tag{2.33}$$

On établit la condition initiale pour $\mathbf{v}_2$ à l'aide de la formule (2.29). Puisque $\mathbf{v}_1(0) = 0$, on a

$$\mathbf{v}_2(0) + \tau\boldsymbol{\eta}^\tau(0) = 0.$$

D'où $\mathbf{v}_2(0) = 0$. Cette condition initiale, l'équation (2.33) et la première équation (2.26) forment le problème

$$\begin{gathered}\frac{d\mathbf{v}_2}{dt} + A_1\mathbf{w}_2 + A_2\mathbf{v}_2 = -\frac{1}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{1}{2}\frac{d^2\mathbf{v}_1}{dt^2}, \qquad t\in(0, T], \\ \mathbf{w}_2 - \mathbf{v}_2 = A_2\mathbf{v}_1, \qquad t\in(0, T], \\ \mathbf{v}_2(0) = 0. \end{gathered} \tag{2.34}$$

On suppose qu'il y a existence et que la fonction $\mathbf{v}_2$ est suffisamment régulière. On établit l'unicité comme pour le système (2.30). Ainsi, la construction de $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{w}_1$, $\mathbf{w}_2$ est terminée. On passe au système (2.31), (2.32). On adjoint à ce système la condition initiale homogène résultant de l'équation (2.29) et des conditions $\mathbf{v}_1(0) = \mathbf{v}_2(0) = 0$, il vient

$$\begin{gathered}\frac{\eta^\tau - \zeta^\tau}{\tau} + A_2\eta^\tau = \mathbf{g}_1 \equiv \frac{1}{\tau}\psi_1, \qquad t\in\omega_\tau, \\ \frac{\zeta^\tau - \eta^\tau(t-\tau)}{\tau} + A_1\zeta^\tau = \mathbf{g}_2 \equiv \frac{1}{\tau}\psi_2 + \psi_3, \qquad t\in\omega_\tau, \\ \eta^\tau(0) = 0. \end{gathered} \tag{2.35}$$

Toutes les quantités ψ_i sont bornées:

$$\|\psi_i\| \leqslant c_3, \qquad t\in\omega_\tau, \qquad i = 1, 2, 3. \tag{2.36}$$

On montre la propriété de borne de la solution du problème (2.35), bien que le second membre soit de l'ordre de $1/\tau$. Soit le problème

$$\begin{gathered}\frac{\rho^\tau - \sigma^\tau}{\tau} = \frac{1}{\tau}\psi_1, \qquad t\in\omega_\tau, \\ \frac{\sigma^\tau - \rho^\tau(t-\tau)}{\tau} = \frac{1}{\tau}\psi_2, \qquad t\in\omega_\tau, \\ \rho^\tau(0) = 0. \end{gathered} \tag{2.37}$$

Sa solution est cherchée sous forme explicite:

$$\begin{gathered}\rho^\tau = 0, \qquad t\in\bar\omega_\tau \\ \sigma^\tau = \psi_2, \qquad t\in\omega_\tau. \end{gathered} \tag{2.38}$$

On a utilisé essentiellement le fait que

$$\psi_1 + \psi_2 = 0, \qquad t\in\omega_\tau,$$

en vertu de (2.33). On modifie le problème (2.37) comme suit:

$$\begin{gathered}\frac{\rho^\tau - \sigma^\tau}{\tau} + A_2\rho^\tau = \frac{1}{\tau}\psi_1, \qquad t\in\omega_\tau, \\ \frac{\sigma^\tau - \rho^\tau(t-\tau)}{\tau} + A_1\sigma^\tau = \frac{1}{\tau}\psi_2 + A_1\psi_2, \qquad t\in\omega_\tau, \\ \rho^\tau(0) = 0. \end{gathered} \tag{2.39}$$

On pose $\eta^\tau - \rho^\tau = \varepsilon^\tau$, $\zeta^\tau - \sigma^\tau = \delta^\tau$ et on retranche les équations (2.39) des équations (2.35) correspondantes, il vient le problème

$$\frac{\varepsilon^\tau - \delta^\tau}{\tau} + A_2 \varepsilon^\tau = 0,$$

$$\frac{\delta^\tau - \varepsilon^\tau (t - \tau)}{\tau} + A_1 \delta^\tau = \psi_3 - A_1 \psi_2, \tag{2.40}$$

$$\varepsilon^\tau (0) = 0.$$

La fonction ψ_2 étant connue, on démontre aisément la propriété de borne de $A_1 \psi_2$. Aussi le second membre dans le dernier problème est borné, et l'estimation (2.5) conduit à

$$\max_{\bar\omega_\tau} \| \varepsilon^\tau \| \leqslant c_4, \qquad \max_{\omega_\tau} \| \delta^\tau \| \leqslant c_4.$$

Vu que

$$\eta^\tau = \rho^\tau + \varepsilon^\tau, \qquad \zeta^\tau = \sigma^\tau + \delta^\tau.$$

on a

$$\max_{\bar\omega_\tau} \| \eta^\tau \| \leqslant \max_{\bar\omega_\tau} \| \rho^\tau \| + \max_{\bar\omega_\tau} \| \varepsilon^\tau \| \leqslant c_4.$$

$$\max_{\omega_\tau} \| \zeta^\tau \| \leqslant \max_{\omega_\tau} \| \sigma^\tau \| + \max_{\omega_\tau} \| \delta^\tau \| \leqslant c_3 + c_4.$$

Ainsi, le développement (2.19) existe. Théoriquement, on procède de même en ce qui concerne le développement (2.20). On ne saurait toutefois oublier que $\mathbf{w}_1$ et $\mathbf{w}_2$ sont en général moins régulières que $\mathbf{v}_1$ et $\mathbf{v}_2$ respectivement et qu'on les trouve avec une précision plus mauvaise. Cela signifie que la précision de la solution corrigée du cas (2.19) est supérieure à celle du cas (2.20). La différence est particulièrement frappante au voisinage des frontières des domaines géométriques pour les équations aux dérivées partielles.

5.3. Equation de la chaleur en dimension deux

Nous allons illustrer la méthode décrite dans ses grandes lignes dans le paragraphe précédent sur l'exemple d'équation de la chaleur.

Soit Ω un domaine borné bidimensionnel de frontière Γ. On désigne par Q le cylindre ouvert $\Omega \times (0, T)$ de surface latérale $S = \Gamma \times [0, T]$ et on considère l'équation

$$\frac{\partial u}{\partial t} = \Delta u + f \quad \text{dans } Q \tag{3.1}$$

avec les conditions initiales et aux limites

$$u(\mathbf{x}, 0) = 0, \qquad \mathbf{x} \in \Omega, \tag{3.2}$$

$$u(\mathbf{x}, t) = 0, \qquad (\mathbf{x}, t) \in S. \tag{3.3}$$

Nous avons introduit en dimension un les classes de régularité caractéristiques de la solution du problème. En dimension deux, on suit [25]. On désigne par $H^l(Q)$, l non entier quelconque, l'espace de Banach des fonctions $v(\mathbf{x}, t)$ continues sur Q, ainsi que leurs dérivées

$$\frac{\partial^{r+s_1+s_2} v}{\partial t^r \partial x^{s_1} \partial y^{s_2}} \qquad \text{pour } 2r + s_1 + s_2 < l$$

muni de la norme finie

$$\| u \|_{H^l(\bar{Q})} = \langle u \rangle_Q^{(l)} + \sum_{j=0}^{[l]} \langle u \rangle_Q^{(j)}, \tag{3.4}$$

où

$$\langle u \rangle_Q^{(j)} = \sum_{2r+s_1+s_2=j} \max_{\bar{Q}} \left| \frac{\partial^{r+s_1+s_2} u}{\partial t^r \partial x^{s_1} \partial y^{s_2}} \right|, \quad j = 0, 1, \ldots, [l].$$

$$\langle u \rangle_Q^{(l)} = \langle u \rangle_x^{(l)} + \langle u \rangle_t^{(l/2)},$$

$$\langle u \rangle_x^{(l)} = \sum_{2r+s_1+s_2=[l]} \left\langle \frac{\partial^{r+s_1+s_2} u}{\partial t^r \partial x^{s_1} \partial y^{s_2}} \right\rangle_x^{l-[l]},$$

$$\langle u \rangle_t^{(l/2)} = \sum_{0<l-2r-s_1-s_2<2} \left\langle \frac{\partial^{r+s_1+s_2} u}{\partial t^r \partial x^{s_1} \partial y^{s_2}} \right\rangle_t^{\frac{l-2r-s_1-s_2}{2}}$$

Les quantités $\langle u \rangle_x^{\alpha}$ et $\langle u \rangle_t^{\beta}$ ont pour $\alpha, \beta \in (0, 1)$ le même sens que dans § 5.1.

Les conditions initiales et aux limites étant homogènes, la *condition de concordance d'ordre* 0 a lieu automatiquement: si $(x, y) \in \Gamma$, alors

$$\lim_{t \to 0} u(x, y, t) = \lim_{(x', y') \to (x, y)} u(x', y', 0). \tag{3.5}$$

La *condition d'ordre* 1 nécessaire pour la solution ayant les dérivées partielles $\partial u/\partial t$, $\partial^2 u/\partial x^2$, $\partial^2 u/\partial y^2$ continues sur $\bar{Q}$ s'écrit

$$f(x, y, 0) = 0 \qquad \forall (x, y) \in \Gamma. \tag{3.6}$$

La dérivée seconde $\partial^2 u/\partial t^2$ est nulle sur S car la condition aux limites (3.3) est homogène. On la cherche à l'aide de l'équation (3.1):

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}(\Delta u + f) = \Delta\left(\frac{\partial u}{\partial t}\right) + \frac{\partial f}{\partial t} = \Delta(\Delta u + f) + \frac{\partial f}{\partial t} =$$
$$= \Delta \Delta u + \Delta f + \frac{\partial f}{\partial t}.$$

Le terme $\Delta \Delta u$ est nul pour $t = 0$ (en vertu de la condition initiale (3.2)), si bien que la *condition de concordance d'ordre* 2

$$\Delta f(\mathbf{x}, 0) + \frac{\partial f}{\partial t}(\mathbf{x}, 0) = 0 \qquad \forall \mathbf{x} \in \Gamma \tag{3.7}$$

est nécessaire pour l'existence des dérivées continues $\partial^2 u/\partial t^2$ et $\Delta\Delta u$. Les conditions d'ordre supérieur sont formées de façon analogue (voir [25]). Comme dans le cas unidimensionnel, ces conditions de concordance sont non seulement nécessaires, mais aussi suffisantes pour que les dérivées concernées soient continues sur $\overline{Q}$ pour le second membre f régulier.

THÉORÈME 3.1 (voir [25]). *Soit $l > 0$ non entier, $f \in H^l(\overline{Q})$, $\Gamma \in C^{l+2}$. Si l'on est dans les conditions de concordance d'ordre 0, 1, ..., $[l/2]+1$, le problème* (3.1) *à* (3.3) *possède une solution unique $u \in H^{l+2}(\overline{Q})$.*

On construit le schéma de décomposition. S'agissant des schémas localement de dimension un [43], l'extrapolation de Richardson fournit une solution approchée exacte à l'ordre $\tau^2 + h^3$ parce que dans le cas d'analogues à trois points de la dérivée seconde on ne réussit pas à compromettre l'influence de l'erreur irrégulière commise au voisinage de la frontière. Ce résultat sera énoncé de façon rigoureuse à la fin du paragraphe. Son intérêt tient à son rôle dans le raffinement d'un schéma aux différences d'usage courant.

Le problème de diminuer la contribution de l'erreur irrégulière limitrophe a été discuté plus haut (voir § 4.2), et on l'a abordé de deux manières différentes. Dans le premier cas, on a réalisé une approximation spéciale à plusieurs points des conditions aux limites au voisinage de la frontière, et, dans le second, on a procédé de même en ce qui concerne les dérivées secondes, l'approximation étant basée sur des nœuds irrégulièrement espacés. Chaque fois, on n'a utilisé à l'intérieur du domaine que le schéma standard (à cinq points). La seconde tactique s'avère plus avantageuse pour les équations paraboliques (en tout cas pour obtenir une précision en h^3 et h^4).

On discrétise d'abord par rapport à l'espace. On suppose que le domaine Ω est inclus dans le carré $\{-b < x < b,\ -b < y < b\}$ et on couvre ce carré d'un réseau rectangulaire régulier de pas $h = b/N$ formé de lignes $x_i = ih$, $y_j = jh$, $i, j = -N, \dots, N$. On désigne par Ω_h l'ensemble des nœuds intérieurs à Ω. Conformément aux notations des n°s 4.2.1 et 4.2.3, on introduit les ensembles $\Gamma_{h,x}$, $\Omega^r_{h,x}$, $\Omega^{ir}_{h,x}$, $\Gamma_{h,y}$, $\Omega^r_{h,y}$ et $\Omega^{ir}_{h,y}$. On approche les opérateurs $L_1 = \partial^2/\partial x^2$ et $L_2 = \partial^2/\partial y^2$ moyennant une formule à trois points ou une formule à quatre points selon qu'il s'agit des nœuds réguliers ou irréguliers. Soit (x, y) un nœud de Ω_h. Quatre cas peuvent se présenter:

a) Le nœud (x, y) est régulier dans la direction x. On a

$$L_1^h u(x, y) \equiv u_{\bar{x}x}(x, y) = \frac{u(x-h, y) - 2u(x, y) + u(x+h, y)}{h^2}. \tag{3.8}$$

b) Le nœud (x, y) est irrégulier dans la direction x. On trouve dans la direction parallèle à l'axe Ox un point $(\xi, y) \in \Gamma_{h,x}$ qui est le plus proche de (x, y) et tel que la distance de (ξ, y) et (x, y) soit inférieure à h. Supposons cette distance être égale à δh, où $\delta \in (0, 1)$. On pose

$$L_1^h u(x, y) = \frac{6}{\delta(\delta+1)(\delta+2)h^2} u(\xi, y) - \frac{3-\delta}{\delta h^2} u(x, y) +$$

$$+ \frac{4-2\delta}{(1+\delta)h^2} u(x \pm h, y) - \frac{1-\delta}{(2+\delta)h^2} u(x \pm 2h, y). \quad (3.9)$$

Si $\xi > x$ (voir fig. 4.2), on prend le signe moins. Dans le cas contraire, on choisit le signe plus. Il se peut que le point $(x \pm 2h, y)$ soit extérieur au domaine, auquel cas l'approximation n'a pas de sens. La remarque du n° 4.2.3 indique un moyen de sortir d'affaire. On transforme certains points réguliers en des points irréguliers. On n'oubliera pas qu'on a des fois l'inégalité $1 < \delta < 3/2$. On note que δ est défini en chaque point de $\Omega_{h,x}^{ir}$, i.e. on a construit une fonction positive $\delta(\mathbf{x})$ de domaine de définitions $\Omega_{h,x}^{ir}$.

c) Le nœud (x, y) est régulier dans la direction y. On a

$$L_2^h u(x, y) \equiv u_{\bar{y}y}(x, y) = \frac{u(x, y-h) - 2u(x, y) + u(x, y+h)}{h^2}. \quad (3.10)$$

d) Le nœud (x, y) est irrégulier dans la direction y. Il existe dans la direction parallèle à l'axe Oy un point $(x, \eta) \in \Gamma_{h,y}$ qui est le plus proche de (x, y). La distance correspondante est ρh, avec $\rho \in (0, 1)$. On pose

$$L_2^h u(x, y) = \frac{6}{\rho(\rho+1)(\rho+2)h^2} u(x, \eta) - \frac{3-\rho}{\rho h^2} u(x, y) +$$

$$+ \frac{4-2\rho}{(1+\rho)h^2} u(x, y \pm h) - \frac{1-\rho}{(2+\rho)h^2} u(x, y \pm 2h). \quad (3.11)$$

Le signe moins correspond à $\eta > y$. Si $\eta < y$, on prend le signe plus. Il se peut que l'intervalle d'extrémités (x, η) et $(x, y \pm 2h)$ n'appartienne pas à $\bar{\Omega}$; nous renvoyons le lecteur une fois de plus à la remarque du n° 4.2.3 et nous rappelons qu'avec le procédé mentionné, la quantité ρ peut être pour certains nœuds $\mathbf{x} \in \Omega_{h,y}^{ir}$ supérieure à 1, mais inférieure à 3/2.

Ainsi, on vient de construire une fonction $\rho(\mathbf{x})$ définie en chaque point $\mathbf{x} \in \Omega_{h,y}^{ir}$.

Voyons le schéma de décomposition. Sur l'intervalle $[0, T]$, on établit le réseau

$$\bar{\omega}_\tau = \{t_k = k\tau;\quad k = 0, 1, \ldots, M\}$$

de pas $\tau = T/M$. On pose

$$\omega_\tau = \{t_k \in \bar{\omega}_\tau,\ k \neq 0\}.$$

On définit le réseau dans le cylindre $Q = \Omega \times (0, T)$ comme produit cartésien $Q_h^\tau = \Omega_h \times \omega_\tau$ et on introduit les notations

$$\bar{Q}_h^\tau = \Omega_h \times \bar{\omega}_\tau, \qquad S_h^\tau = \Gamma_h \times \bar{\omega}_\tau.$$

Nous allons nous servir d'un schéma aux différences formé de problèmes à une variable d'espace répétitifs. On omet la variable indépendante $(\mathbf{x}, t) = (x, y, t)$ et on écrit

$$\frac{u^* - u^\tau(x, y, t-\tau)}{\tau} - L_1^h u^* = f, \qquad (x, y, t) \in Q_h^\tau,$$

$$u^* = 0, \qquad (x, y, t) \in \Gamma_{h,x} \times \omega_\tau, \tag{3.12}$$

$$\frac{u^\tau - u^*}{\tau} - L_2^h u^\tau = 0, \qquad (x, y, t) \in Q_h^\tau,$$

$$u^\tau = 0, \qquad (x, y, t) \in \Gamma_{h,y} \times \omega_\tau. \tag{3.13}$$

Le premier pas est soumis à la condition

$$u^\tau(x, y, 0) = 0, \qquad (x, y) \in \Omega_h. \tag{3.14}$$

Voici un procédé pour chercher la solution du problème discret (3.12) à (3.14). Connaissant $u^\tau(x, y, t-\tau)$ pour tout (x, y) de Ω_h, on trouve à chaque niveau temporel les valeurs $u^*(x, y, t)$ $\forall (x, y) \in \Omega_h$ par le système (3.12), puis toutes les valeurs $u^\tau(x, y, t)$ aux nœuds $(x, y) \in \Omega_h$ moyennant (3.13). On initialise au pas de temps suivant avec $u^\tau(x, y, t)$ ainsi obtenues.

Si l'on interprète (3.12) comme système par rapport aux inconnues $u^*(x, y, t)$, il est un ensemble de sous-systèmes « presque » tridiagonaux. En effet, la matrice serait tridiagonale s'il n'y avait deux éléments « superflus » par sous-système:

$$\begin{bmatrix} b_1 & c_1 & d_1 & & & & \\ a_2 & b_2 & c_2 & & & & \\ & a_3 & \cdot & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & c_{n-2} & \\ & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & d_n & a_n & b_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \cdot \\ \cdot \\ \cdot \\ g_{n-1} \\ g_n \end{bmatrix}. \tag{3.15}$$

Ici z_i sont les valeurs $u^*(x, y, t)$ prises dans certain ordre :

$$z_i = u^*(x_{l+i}, y, t), \qquad i = 1, \dots, n.$$

et le second membre contient les valeurs de f et u^τ obtenues au pas de temps précédent :

$$g_i = f(x_{l+i}, y, t) + \frac{1}{\tau} u^\tau (x_{l+i}, y, t - \tau).$$

Seules figurent dans (3.15) les valeurs inconnues u^* dont les variables indépendantes (x, y) sont des nœuds de Ω_h pour un certain y fixé, et l'ensemble de ces nœuds est tel que (x_l, y) et (x_{l+n}, y) soient irréguliers dans la direction x et que tous les nœuds intermédiaires soient réguliers. Il est clair que s'agissant de Ω non convexe, ces systèmes sont plusieurs pour y fixé.

On introduit les notations $\sigma^- = \delta(x_l, y)$ et $\sigma^+ = \delta(x_{l+n}, y)$. Les coefficients du système (3.15) se récrivent

$$b_1 = \frac{1}{\tau} + \frac{3 - \sigma^-}{\sigma - h^2}, \qquad e_1 = -\frac{4 - \sigma^-}{(1 + \sigma^-) h^2}, \qquad d_1 = \frac{1 - \sigma^-}{(2 + \sigma^-) h^2},$$

$$a_i = e_i = -\frac{1}{h^2}, \qquad b_i = \frac{1}{\tau} + \frac{2}{h^2}, \qquad i = 2, \ldots, n-1, \qquad (3.16)$$

$$d_n = \frac{1 - \sigma^+}{(2 + \sigma^+) h^2}, \qquad a_n = -\frac{4 - 2\sigma^+}{(1 + \sigma^+) h^2}, \qquad b_n = \frac{1}{\tau} + \frac{3 - \sigma^+}{\sigma^+ h^2}.$$

On démontre que quel que soit le rapport de τ et h, le système (3.15) est à dominance diagonale stricte (voir [119]). Deux cas peuvent se présenter pour la première équation :

a) $\sigma^- \in (0, 1]$. On a

$$|b_1| - |e_1| - |d_1| = \frac{1}{\tau} + \frac{3 - \sigma^-}{\sigma^- h^2} - \frac{4 - 2\sigma^-}{(1 + \sigma^-) h^2} - \frac{1 - \sigma^-}{(2 + \sigma^-) h^2} \geqslant$$

$$\geqslant \frac{1}{\tau} + \frac{6 - 5\sigma^- + 2(\sigma^-)^2}{2\sigma^- (1 + \sigma^-) h^2} \geqslant \frac{1}{\tau} + \frac{3}{4\sigma^- h^2} > \frac{1}{\tau} + \frac{1}{2\sigma^- h^2}.$$

b) $\sigma^- \in (1, 3/2]$. Alors

$$|b_1| - |e_1| - |d_1| = \frac{1}{\tau} + \frac{3 - \sigma^-}{\sigma^- h^2} - \frac{4 - 2\sigma^-}{(1 - \sigma^-) h^2} + \frac{1 - \sigma^-}{(2 + \sigma^-) h^2} \geqslant$$

$$\geqslant \frac{1}{\tau} + \frac{9 - 5\sigma^- + 2(\sigma^-)^2}{3\sigma^- (1 + \sigma^-) h^2} \geqslant \frac{1}{\tau} + \frac{47}{60\sigma^- h^2} > \frac{1}{\tau} + \frac{1}{2\sigma^- h^2}.$$

Ainsi,

$$|b_1| - |e_1| - |d_1| \geqslant \frac{1}{\tau} + \frac{1}{2h^2\sigma^-} \qquad (3.17)$$

dans les deux cas. On prouve de même que

$$|b_n| - |a_n| - |d_n| \geqslant \frac{1}{\tau} + \frac{1}{2h^2\sigma^+} \qquad (3.18)$$

et que

$$|b_i| - |a_i| - |c_i| \geq \frac{1}{\tau} \tag{3.19}$$

pour i restants.

Ces inégalités sont suffisantes pour que le système (3.15) soit non dégénéré (voir [119]). Si l'on élimine au préalable la première et la dernière inconnue, le système est résolu par balayage.

Il en est de même de (3.13). En effet, (3.13) considéré comme système par rapport aux valeurs inconnues $u^\tau(x, y, t)$ est un ensemble de systèmes « presque » tridiagonaux de la forme (3.15). Cette fois z_i sont les valeurs $u^\tau(x, y, t)$:

$$z_i = u^\tau(x, y_{m+i}, t), \qquad i = 1, \ldots, q,$$

et le second membre renferme u^* obtenus antérieurement:

$$g_i = \frac{1}{\tau} u^*(x, y_{m+i}, t).$$

Chaque système partiel (3.15) contient les inconnues u^τ dont les variables indépendantes (x, y) constituent des nœuds de Ω_h pour un certain $\bar{x}$ fixé, et si Ω n'est pas convexe, il correspond à un seul x plusieurs sous-systèmes. Les inégalités (3.17) à (3.19) ont lieu cette fois encore. Le résultat de non-dégénérescence reste en vigueur, et on peut recourir au balayage quitte à effectuer certaines transformations élémentaires.

Tout ce que nous savons sur la possibilité du problème (3.12) à (3.14) autorise à conclure à l'existence d'au moins une solution et suggère en même temps un procédé de recherche économique de celle-ci. Le passage d'un niveau temporel à un autre exige des opérations arithmétiques en nombre proportionnel au nombre de nœuds du réseau Ω_h.

On se propose de démontrer une estimation à priori caractéristique de la stabilité numérique du passage indiqué.

THÉORÈME 3.2. *La solution du problème*

$$\begin{aligned} &\frac{v^* - v^\tau(x, y, t-\tau)}{\tau} - L_1^h v^* = f_1, \qquad (x, y, t) \in Q_h^\tau, \\ &\qquad\qquad v^* = 0, \qquad (x, y, t) \in \Gamma_{h,x} \times \omega_\tau, \end{aligned} \tag{3.20}$$

$$\begin{aligned} &\frac{v^\tau - v^*}{\tau} - L_2^h v^\tau = f_2, \qquad (x, y, t) \in Q_h^\tau, \\ &\qquad\qquad v^\tau = 0, \qquad (x, y, t) \in \Gamma_{h,y} \times \omega_\tau, \end{aligned} \tag{3.21}$$

avec la condition initiale

$$v^\tau(x, y, 0) = 0, \qquad (x, y) \in \Omega_h. \tag{3.22}$$

vérifie l'inégalité

$$\max_{Q_h^\tau} |v^\tau| \leqslant \sum_{t\in\omega_\tau} \tau \{\max_{\Omega_{h,x}^r} |f_1| + \max_{\Omega_{h,y}^r} |f_2|\} +$$
$$+ 2h^2 \{\max_{\Omega_{h,x}^{ir}\times\omega_\tau} |\delta f_1| + \max_{\Omega_{h,y}^{ir}\times\omega_\tau} |\rho f_2|\}.$$

DÉMONSTRATION. On met les fonctions f_1 et f_2 associées aux nœuds du réseau de discrétisation Q_h^τ sous forme de somme:

$$f_1 = f_1^r + f_1^{ir}, \qquad f_2 = f_2^r + f_2^{ir}.$$

Le support de f_1^r est concentré aux nœuds du domaine $\Omega_{h,x}^r \times \omega_\tau$ et celui de f_1^{ir} aux nœuds de $\Omega_{h,x}^{ir} \times \omega_\tau$:

$$f_1^r = \begin{cases} f_1 \text{ sur } \Omega_{h,x}^r \times \omega_\tau \\ 0 \text{ sur } \Omega_{h,x}^{ir} \times \omega_\tau, \end{cases} \qquad f_1^{ir} = \begin{cases} 0 \text{ sur } \Omega_{h,x}^r \times \omega_\tau, \\ f_1 \text{ sur } \Omega_{h,x}^{ir} \times \omega_\tau. \end{cases}$$

On a de même pour f_2:

$$f_2^r = \begin{cases} f_2 \text{ sur } \Omega_{h,y}^r \times \omega_\tau, \\ 0 \text{ sur } \Omega_{h,y}^{ir} \times \omega_\tau, \end{cases} \qquad f_2^{ir} = \begin{cases} 0 \text{ sur } \Omega_{h,y}^r \times \omega_\tau, \\ f_2 \text{ sur } \Omega_{h,y}^{ir} \times \omega_\tau. \end{cases}$$

Etant donnée cette décomposition, on pose

$$v^\tau = w_1^\tau + w_2^\tau, \qquad v^* = w_1^* + w_2^*,$$

avec w_1^*, w_1^τ solutions du problème (3.20) à (3.22) pour f_1^r et f_2^r respectivement, et w_2^*, w_2^τ solutions de ce problème pour les seconds membres respectifs f_1^{ir} et f_2^{ir}. Il s'agit d'évaluer les solutions. Plaçons-nous dans le cas w_1^*, w_1^τ.

On suppose qu'on a l'inégalité

$$\max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t-\tau)| \leqslant \sum_{\substack{t'\in\omega_\tau \\ t'\leqslant t-\tau}} \tau (\max_{Q_h^\tau} |f_1^r| + \max_{Q_h^\tau} |f_2^r|). \qquad (3.23)$$

Prouvons-la pour le niveau de temps t. Voyons les solutions de (3.15) pour deux types de second membre. On pose

$$g_i = \frac{1}{\tau} w_1^\tau(x_{l+i}, y, t)$$

(f_1^r est nul), auquel cas

$$\max_{(x,y)\in\Omega_h} |w_1^*(x, y, t)| \leqslant \max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t-\tau)|. \qquad (3.24)$$

On raisonne par l'absurde et on suppose que (3.24) n'est pas juste. Alors $|w_1^{\bullet}(x, y, t)|$ atteint en un nœud $(x_0, y_0) \in \Omega_h$ son maximum, et

$$|w_1^{\bullet}(x_0, y_0, t)| > \max_{(x,y)\in\Omega_h} |w_1^{\tau}(x, y, t-\tau)|. \tag{3.25}$$

Quelle est l'équation associée? Si (x_0, y_0) est régulier dans la direction x, on a

$$\left(\frac{2}{h^2}+\frac{1}{\tau}\right) w_1^{\bullet}(x_0, y_0, t) =$$

$$= \frac{1}{h^2} w_1^{\bullet}(x_0 - h, y_0, t) + \frac{1}{h^2} w_1^{\bullet}(x_0 + h, y_0, t) + \frac{1}{\tau} w_1^{\tau}(x_0, y_0, t-\tau).$$

Cette égalité et (3.25) entraînent

$$\left(\frac{2}{h^2}+\frac{1}{\tau}\right) |w_1^{\bullet}(x_0, y_0, t)| < \left(\frac{2}{h^2}+\frac{1}{\tau}\right) |w_1^{\bullet}(x_0, y_0, t)|:$$

contradiction qui réfute l'hypothèse de (x_0, y_0) régulier dans la direction x.

On suppose que ce nœud est irrégulier dans la direction x, auquel cas on a, en notations de (3.15), (3.16),

$$\left(\frac{1}{\tau}+\frac{3-\sigma^{\pm}}{\sigma^{\pm} h^2}\right) w_1^{\bullet}(x_0, y_0, t) = \frac{4-2\sigma^{\pm}}{(1+\sigma^{\pm})h^2} w_1^{\bullet}(x_0, h \pm y_0, t) -$$

$$- \frac{1-\sigma^{\pm}}{(2+\sigma^{\pm})h^2} w_1^{\bullet}(x_0 \pm 2h, y_0, t) + \frac{1}{\tau} w_1^{\tau}(x_0, y_0, t-\tau).$$

Le signe est fonction de la position du nœud. On passe aux valeurs absolues et on utilise (3.25), il vient

$$\left(\frac{1}{\tau}+\frac{3-\sigma^{\pm}}{\sigma^{\pm} h^2}\right) |w_1^{\bullet}(x_0, y_0, t)| <$$

$$< \left(\frac{4-2\sigma^{\pm}}{(1+\sigma^{\pm})h^2} + \frac{|1-\sigma^{\pm}|}{(2+\sigma^{\pm})h^2} + \frac{1}{\tau}\right) |w_1^{\bullet}(x_0, y_0, t)|.$$

Etant donnés (3.17), (3.18), il y a contradiction:

$$\left(\frac{1}{\tau}+\frac{3-\sigma^{\pm}}{\sigma^{\pm} h^2}\right) |w_1^{\bullet}(x_0, y_0, t)| \leq \left(\frac{1}{\tau}+\frac{3-\sigma^{\pm}}{\sigma^{\pm} h^2}\right) |w_1^{\bullet}(x_0, y_0, t)|.$$

Aussi l'inégalité (3.25) n'est juste en aucun point de Ω_h et on a (3.24) pour le second membre nul.

Soit

$$g_i = f_1^{\tau}(x_{i+1}, y, t)$$

(les données initiales w_1^τ sont nulles). On démontre qu'on a au point de maximum de $|\overset{\bullet}{w}_1 (x, y, t)|$ la majoration

$$|\overset{\bullet}{w}_1 (x_0, y_0, t)| \leqslant \tau |f_1^\tau (x_0, y_0, t)|. \tag{3.26}$$

Si le point (x_0, y_0) est régulier dans la direction x, alors

$$\left(\frac{2}{h^2} + \frac{1}{\tau}\right) \overset{\bullet}{w}_1 (x_0, y_0, t) =$$
$$= \frac{1}{h^2} \overset{\bullet}{w}_1 (x_0 - h, y_0, t) + \frac{1}{h^2} \overset{\bullet}{w}_1 (x_0 + h, y_0, t) + f_1^\tau (x_0, y_0, t).$$

On passe aux valeurs absolues et on remplace la fonction dans le second membre par son majorant:

$$\left(\frac{2}{h^2} + \frac{1}{\tau}\right) |\overset{\bullet}{w}_1 (x_0, y_0, t)| \leqslant \frac{2}{h^2} \overset{\bullet}{w}_1 (x_0, y_0, t)| + |f_1^\tau (x_0, y_0, t)|.$$

D'où (3.26).

Si (x_0, y_0) est irrégulier dans la direction x, alors

$$\left(\frac{1}{\tau} + \frac{3-\sigma^\pm}{\sigma^\pm h^2}\right) \overset{\bullet}{w}_1 (x_0, y_0, t) = \frac{4 - 2\sigma^\pm}{(1+\sigma^\pm) h^2} \overset{\bullet}{w}_1 (x_0 \pm h, y_0, t) -$$
$$- \frac{1-\sigma^\pm}{(2+\sigma^\pm) h^2} \overset{\bullet}{w}_1 (x_0 \pm 2h, y_0, t) + f_1^\tau (x_0, y_0, t).$$

On prend le module des deux membres et on remplace le second membre par son majorant, il vient

$$\left(\frac{1}{\tau} + \frac{3-\sigma^\pm}{\sigma^\pm h^2}\right) |\overset{\bullet}{w}_1 (x_0, y_0, t)| \leqslant$$
$$\leqslant \left(\frac{4 - 2\sigma^\pm}{(1+\sigma^\pm) h^2} + \frac{|1-\sigma^\pm|}{(2+\sigma^\pm) h^2}\right) |\overset{\bullet}{w}_1 (x_0, y_0, t)| + |f_1^\tau (x_0, y_0, t)|.$$

Etant donnés (3.17), (3.18), on obtient

$$\left(\frac{1}{\tau} + \frac{3-\sigma^\pm}{\sigma^\pm h^2}\right) |\overset{\bullet}{w}_1 (x_0, y_0, t)| \leqslant \frac{3-\sigma^\pm}{\sigma^\pm h^2} |\overset{\bullet}{w}_1 (x_0, y_0, t) + | f_1^\tau (x_0, y_0, t)|,$$

d'où l'estimation (3.26). Comme (x_0, y_0) réalise le maximum de $|\overset{\bullet}{w}_1 (x, y, t)|$ pour t fixé, on a

$$\max_{(x,y)\in\Omega_h} |\overset{\bullet}{w}_1 (x, y, t)| \leqslant \tau \max_{(x,y)\in\Omega_h} |f_1^\tau (x, y, t)|. \tag{3.27}$$

Soit le cas où le second membre et les données initiales sont non nuls. La solution $\overset{\bullet}{w}_1$ du problème

$$\frac{\overset{\bullet}{w}_1 - w_1^\tau(x, y, t-\tau)}{\tau} - L_1^h \overset{\bullet}{w}_1 = f_1^\tau, \qquad (x, y) \in \Omega_h,$$
$$\overset{\bullet}{w}_1 = 0, \qquad (x, y) \in \Gamma_{h,x},$$

admet par suite de (3.24) et (3.27) la majoration

$$\max_{(x,y)\in\Omega_h} |\overset{\bullet}{w}_1(x, y, t)| \leqslant$$

$$\leqslant \max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t-\tau)| + \tau \max_{(x,y)\in\Omega_h} |f_1^\tau(x, y, t)|. \qquad (3.28)$$

On démontre de même pour la solution w_1^τ du problème

$$\frac{w_1^\tau - \overset{\bullet}{w}_1}{\tau} - L_2^h w_1^\tau = f_2^\tau, \qquad (x, y) \in \Omega_h, \qquad w_1^\tau = 0, \qquad (x, y) \in \Gamma_{h,y},$$

l'inégalité

$$\max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t)| \leqslant$$

$$\leqslant \max_{(x,y)\in\Omega_h} |\overset{\bullet}{w}_1(x, y, t)| + \tau \max_{(x,y)\in\Omega_h} |f_2^\tau(x, y, t)|. \qquad (3.29)$$

S'agissant du pas τ entier (lorsqu'on calcule $w_1^\tau(x, y, t)$ à partir de $w_1^\tau(x, y, t-\tau)$), les estimations (3.28), (3.29) ont pour conséquence

$$\max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t)| \leqslant$$

$$\leqslant \max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t-\tau)| + \tau \max_{Q_h^\tau} |f_1^\tau| + \tau \max_{Q_h^\tau} |f_2^\tau|.$$

Si l'on est dans la condition (3.23) pour les niveaux temporels $t' < t$, alors

$$\max_{(x,y)\in\Omega_h} |w_1^\tau(x, y, t)| \leqslant \sum_{\substack{t'\in\omega_\tau \\ t'<t}} \tau \left(\max_{Q_h^\tau} |f_1^\tau| + \max_{Q_h^\tau} |f_2^\tau|\right). \qquad (3.30)$$

Ainsi, on a établi (3.23) pour tous les niveaux de temps.

Soit le problème pour la fonction w_2^τ:

$$\left.\begin{aligned} \frac{\overset{\bullet}{w}_2 - w_2^\tau(x, y, t-\tau)}{\tau} - L_1^h \overset{\bullet}{w}_2 &= f_1^{ir}, & (x, y, t) &\in Q_h^\tau, \\ \overset{\bullet}{w}_2 &= 0, & (x, y, t) &\in \Gamma_{h,y} \times \omega_\tau, \end{aligned}\right\} \qquad (3.31)$$

$$\left.\begin{aligned} \frac{w_2^\tau - \overset{\bullet}{w}_2}{\tau} - L_2^h w_2^\tau &= f_2^{ir}, & (x, y, t) &\in Q_h^\tau, \\ w_2^\tau &= 0, & (x, y, t) &\in \Gamma_{h,y} \times \omega_\tau. \end{aligned}\right\} \qquad (3.32)$$

$$w_2^\tau(x, y, 0) = 0, \qquad (x, y) \in \Omega_h. \qquad (3.33)$$

On démontre l'inégalité

$$\max\{\max_{Q_h^\tau} |w_2^\tau|,\ \max_{Q_h^\tau} |\overset{\bullet}{w}_2|\} \leqslant$$

$$\leqslant 2h^2\left(\max_{\Omega_{h,x}^{ir}\times\omega_\tau} |\delta f_1^{ir}| + \max_{\Omega_{h,y}^{ir}\times\omega_\tau} |\rho f_2^{ir}|\right). \quad (3.34)$$

On suppose que

$$|w_2^\tau(x_0, y_0, t_0)| = \max\{\max_{Q_h^\tau} |w_2^\tau|,\ \max_{Q_h^\tau} |\overset{\bullet}{w}_2|\}. \quad (3.35)$$

Si le point (x_0, y_0) est irrégulier dans la direction y, l'équation associée du système (3.32) s'écrit

$$\left(\frac{3-\rho}{\rho h^2} + \frac{1}{\tau}\right) w_2^\tau(x_0, y_0, t_0) = \frac{4-2\rho}{(1+\rho)h^2} w_2^\tau(x_0, y_0 \pm h, t_0) -$$

$$-\frac{1-\rho}{(2+\rho)h^2} w_2^\tau(x_0, y_0 \pm 2h, t_0) + \frac{1}{\tau} \overset{\bullet}{w}_2(x_0, y_0, t_0) +$$

$$+ f_2^{ir}(x_0, y_0, t_0), \qquad \rho = \rho(x_0, y_0).$$

Le choix du signe est fonction de la position du point concerné et n'influe nullement sur les calculs. On passe aux modules et on utilise (3.35):

$$\left(\frac{3-\rho}{\rho h^2} + \frac{1}{\tau}\right) |w_2^\tau(x_0, y_0, t_0)| \leqslant$$

$$\leqslant \left(\frac{4-2\rho}{(1+\rho)h^2} + \frac{|1-\rho|}{(2+\rho)h^2} + \frac{1}{\tau}\right) |w_2^\tau(x_0, y_0, t_0)| + |f_2^{ir}(x_0, y_0, t_0)|.$$

On pose $\sigma^- = \rho$ ou $\sigma^+ = \rho$ dans les formules (3.16), et les inégalités (3.17), (3.18) s'écrivent

$$\frac{1}{\tau} + \frac{3-\rho}{\rho h^2} - \frac{4-2\rho}{(1+\rho)h^2} - \frac{|1-\rho|}{(2+\rho)h^2} \geqslant \frac{1}{\tau} + \frac{1}{2\rho h^2}.$$

L'inégalité précédente prend maintenant la forme

$$\frac{1}{2\rho h^2} |w_2^\tau(x_0, y_0, t_0)| \leqslant |f_2^{ir}(x_0, y_0, t_0)|.$$

Avec l'hypothèse (3.35), cette estimation entraîne (3.34).

Si le point (x_0, y_0) est régulier dans la direction x, l'équation correspondante du système (3.32) est

$$\left(\frac{2}{h^2} + \frac{1}{\tau}\right) w_2^\tau(x_0, y_0, t_0) = \frac{1}{h^2} w_2^\tau(x_0, y_0 + h, t_0) +$$

$$+ \frac{1}{h^2} w_2^\tau(x_0, y_0 - h, t_0) + \frac{1}{\tau} \overset{\bullet}{w}_2(x_0, y_0, t_0).$$

On passe aux modules. On a compte tenu de (3.35)

$$\left(\frac{2}{h^2} + \frac{1}{\tau}\right) | w_2^\tau (x_0, y_0, t_0) | \leqslant \left(\frac{2}{h^2} + \frac{1}{\tau}\right) | w_2^\tau (x_0, y_0, t_0) |. \quad (3.36)$$

Si l'on admet que

$$| w_2^\tau (x_0, y_0 + h, t_0) | < | w_2^\tau (x_0, y_0, t_0) |,$$

l'inégalité (3.36) devient stricte. La chose est impossible, si bien que

$$| w_2^\tau (x_0, y_0 + h, t_0) | \geqslant | w_2^\tau (x_0, y_0, t_0) |.$$

Cette inégalité et la relation (3.35) conduisent à

$$| w_2^\tau (x_0, y_0, t_0) | = | w_2^\tau (x_0, y_0 + h, t_0) |.$$

Ainsi, $(x_0, y_0 + h, t_0)$ est un autre point de maximum, d'où la validité pour lui de tous les résultats relatifs à (x_0, y_0, t_0). Cela signifie que, ou bien le point $(x_0, y_0 + h)$ est irrégulier dans la direction y (auquel cas on a l'estimation (3.34)), ou bien ce point est régulier dans la direction y, mais

$$| w_2^\tau (x_0, y_0 + h, t_0) | = | w_2^\tau (x_0, y_0 + 2h, t_0) |.$$

Si l'on poursuit dans cette voie, on obtient l'inégalité (3.34) ou la chaîne d'égalités

$$| w_2^\tau (x_0, y_0, t_0) | = \ldots = | w_2^\tau (x_0, y_0 + kh, t_0) |.$$

Le domaine Ω étant borné, cette suite de nœuds $(x_0, y_0 + kh)$ réguliers dans la direction y aboutit après moins de $2N$ pas à un nœud irrégulier dans la direction y, i. e. on revient au cas précédent.

Le cas de $\dot{w}_2$ est traité de façon analogue à la différence qu'on étudie le système (3.31).

Ainsi, on a démontré l'inégalité (3.34). Comme $v^\tau = w_1^\tau + w_2^\tau$, les estimations (3.30) et (3.34) fournissent l'inégalité

$$\max_{(x,y)\in\Omega_h} | v^\tau (x, y, t) | \leqslant \sum_{\substack{t'\in\omega_\tau \\ t'<t}} \tau \left(\max_{\Omega_h^\tau} | f_1^\tau | + \max_{\Omega_h^\tau} | f_2^\tau | \right) +$$
$$+ 2h^2 \left(\max_{\Omega_{h,x}^{ir}\times\omega_\tau} | \delta f_1^{ir} | + \max_{\Omega_{h,y}^{ir}\times\omega_\tau} | \rho f_2^{ir} | \right)$$

juste pour tout $t \in \omega_\tau$. Si l'on augmente le nombre de termes sous $\sum$, on a l'inégalité du théorème 3.2.

On se propose d'établir les développements suivant les puissances de τ et h des fonctions u^τ et u^*, solution du problème (3.12) à (3.14).

Théorème 3.3. *On suppose que le problème* (3.1) *à* (3.3) *satisfait aux conditions de concordance d'ordre* 1, 2 *et que* $f \in H^{4+\varepsilon}(\overline{Q})$, $\Gamma \in C^{6+\varepsilon}$, *la constante* $\varepsilon \in (0, 1)$. *La solution* u^τ, u^* *de la méthode de décomposition* (3.12) *à* (3.14) *admet les développements*

$$\left.\begin{aligned} u^\tau &= u + h^2 v_1 + \tau v_2 + (h^4 + \tau^2)\eta^\tau, \\ u^* &= u + h^2 w_1 + \tau w_2 + (h^4 + \tau^2)\eta^* \end{aligned}\right\} \quad \text{sur} \quad Q_h^\tau, \tag{3.37}$$

où les fonctions v_1, v_2, w_1, w_2 *sont continues sur* $\overline{Q}$ *et indépendantes de* τ *et* h *et les fonctions discrètes* η^τ *et* η^* *sont bornées* :

$$\max_{Q_h^\tau} |\eta^\tau| \leqslant c_1, \qquad \max_{Q_h^\tau} |\eta^*| \leqslant c_1. \tag{3.38}$$

Démonstration. On prend v_1 et w_1 comme solution du problème

$$\frac{\partial v_1}{\partial t} = \frac{\partial^2 v_1}{\partial y^2} + \frac{\partial^2 w_1}{\partial x^2} + \frac{1}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right) \quad \text{dans} \quad Q, \tag{3.39}$$

$$w_1 = v_1 \quad \text{dans} \quad Q, \tag{3.40}$$

$$v_1 = 0 \quad \text{sur} \quad S, \tag{3.41}$$

$$v_1(\mathbf{x}, 0) = 0, \qquad \mathbf{x} \in \Omega. \tag{3.42}$$

On substitue à w_1 de (3.39) ses valeurs (3.40) :

$$\frac{\partial v_1}{\partial t} = \Delta v_1 + \frac{1}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right). \tag{3.43}$$

On est, pour le problème (3.43), (3.41), (3.42), dans les hypothèses du théorème 3.1, où la constante $l = 1 + \lambda$, $\lambda \in (0,1)$. Il existe donc une solution unique $v_1 \in H^{3+\lambda}(\overline{Q})$. On définit $w_1 \in H^{3+\lambda}(\overline{Q})$ à partir de (3.40). Les fonctions v_1 et w_1 sont évidemment indépendantes de τ et h. On note que les conditions imposées aux données du problème (3.1) à (3.3) pourraient garantir une régularité plus grande de v_1 et u si ce n'était la condition de concordance d'ordre 3 non remplie. Ainsi, la dérivée $\partial v_1/\partial t$ est bornée, et $\partial^2 v_1/\partial t^2$ ne l'est pas. Toutefois les dérivées d'ordre supérieur sont à croissance modérée au voisinage de $t = 0$. On démontre cette propriété à la lumière de plusieurs résultats d'ordre secondaire.

Lemme 3.4. *Hypothèses du théorème* 3.3. *On a*

$$t^\alpha \frac{\partial^3 u}{\partial t^3}, \; t^\alpha \frac{\partial^6 u}{\partial x^6}, \; t^\alpha \frac{\partial^6 u}{\partial y^6} \in H^\varepsilon(\overline{Q}) \qquad \forall \alpha \in (\varepsilon, 1).$$

Démonstration. On pose

$$z(\mathbf{x}, t) = \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) - \Delta f(\mathbf{x}, 0) - \frac{\partial f}{\partial t}(\mathbf{x}, 0)$$

et on calcule

$$\frac{\partial z}{\partial t}(\mathbf{x}, t) - \Delta z(\mathbf{x}, t) = \frac{\partial^2}{\partial t^2}\left(\frac{\partial u}{\partial t}(\mathbf{x}, t)\right) -$$
$$- \Delta \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) + \Delta\Delta f(\mathbf{x}, 0) + \Delta \frac{\partial f}{\partial t}(\mathbf{x}, 0).$$

Toutes les dérivées de cette égalité sont continues dans Q. L'équation (3.1) entraîne

$$\frac{\partial^3 u}{\partial t^3} = \frac{\partial^2}{\partial t^2}(\Delta u + f) \quad \text{dans} \quad Q,$$

donc

$$\frac{\partial z}{\partial t} = \Delta z + g \quad \text{dans} \quad Q, \tag{3.44}$$

où

$$g(\mathbf{x}, t) = \frac{\partial^2 f}{\partial t^2}(\mathbf{x}, t) + \Delta\Delta f(\mathbf{x}, 0) + \Delta \frac{\partial f}{\partial t}(\mathbf{x}, 0).$$

On effectue la dérivation de (3.1):

$$\frac{\partial^2 u}{\partial t^2} = \Delta\Delta u + \Delta f + \frac{\partial f}{\partial t},$$

d'où

$$z(\mathbf{x}, 0) = 0 \qquad \forall \mathbf{x} \in \Omega, \tag{3.45}$$

et la condition de concordance (3.7) implique

$$z = 0 \quad \text{sur} \quad S. \tag{3.46}$$

Comme on est, pour le problème (3.44) à (3.46), dans les conditions de concordance d'ordre 0, tandis que les conditions d'ordre 1 ne sont pas satisfaites, le théorème 3.1 entraîne $z \in H^{1+\lambda}(\overline{Q})$ $\forall \lambda \in (0, 1)$.

On passe dans le problème (3.44) à (3.46) à la nouvelle fonction $w = t^\alpha z$:

$$\begin{aligned} &\frac{\partial w}{\partial t} = \Delta w + \alpha t^{\alpha-1} z + t^\alpha g \quad \text{dans} \quad Q, \\ &w(\mathbf{x}, 0) = 0, \qquad \mathbf{x} \in \Omega, \\ &w = 0 \quad \text{sur} \quad S. \end{aligned} \tag{3.47}$$

On tire de l'égalité (3.45) et de $z \in H^{1+\lambda}(\overline{Q})$

$$\lim_{t \to 0} t^{-\sigma} z(\mathbf{x}, t) = 0 \qquad \forall \sigma \in \left(0, \frac{1+\lambda}{2}\right), \qquad \mathbf{x} \in \overline{\Omega}. \tag{3.48}$$

Aussi $t^{\alpha-1} z \in H^\alpha(\overline{Q})$ pour $\lambda > 1 - \alpha$. Vu que $\alpha > \varepsilon$, on a $t^{\alpha-1} z \in H^\varepsilon(\overline{Q})$. Ainsi, le second membre de l'équation du problème (3.47) appar-

tient à $H^\varepsilon(\overline{Q})$. Ce problème vérifie les conditions de concordance d'ordre 0 et 1, la dernière propriété résultant de l'égalité

$$\lim_{t\to 0}\left(t^\alpha g(\mathbf{x}, t) + \alpha t^{\alpha-1} z(\mathbf{x}, t)\right) = 0 \qquad \forall \mathbf{x} \in \bar{\Omega}.$$

Aussi $t^\alpha z \in H^{2+\varepsilon}(\overline{Q})$ par suite du théorème 3.1. D'où

$$t^\alpha \frac{\partial^2 z}{\partial x^2}, \ t^\alpha \frac{\partial^2 z}{\partial y^2} \in H^\varepsilon(\overline{Q}), \ \alpha t^{\alpha-1} z + t^\alpha \frac{\partial z}{\partial t} \in H^\varepsilon(\overline{Q}),$$

donc

$$t^\alpha \frac{\partial z}{\partial t} \in H^\varepsilon(\overline{Q}).$$

La définition de la fonction z et la condition de régularité de f autorisent à dire que

$$t^\alpha \frac{\partial^3 u}{\partial t^3}, \ t^\alpha \frac{\partial^4 u}{\partial t^2 \partial x^2}, \ t^\alpha \frac{\partial^4 u}{\partial t^2 \partial y^2} \in H^\varepsilon(\overline{Q}). \tag{3.49}$$

On prend pour z la fonction

$$z(\mathbf{x}, t) = \frac{\partial^3 u}{\partial t\, \partial x^2}(\mathbf{x}, t) - \frac{\partial^2 f}{\partial x^2}(\mathbf{x}, 0)$$

et on raisonne comme pour les formules (3.44) à (3.49), il vient

$$t^\alpha \frac{\partial^5 u}{\partial t\, \partial x^4}, \ t^\alpha \frac{\partial^5 u}{\partial t\, \partial x^2\, \partial y^2} \in H^\varepsilon(\overline{Q}). \tag{3.50}$$

Si on raisonne non sur x mais sur y, alors

$$t^\alpha \frac{\partial^5 u}{\partial t\, \partial y^4} \in H^\varepsilon(\overline{Q}). \tag{3.51}$$

On pose $z = \partial^4 u/\partial x^4$ et on procède de même. Alors

$$t^\alpha \frac{\partial^6 u}{\partial x^6}, \ t^\alpha \frac{\partial^6 u}{\partial x^4\, \partial y^2} \in H^\varepsilon(\overline{Q}) \tag{3.52}$$

et, en remplaçant x par y:

$$t^\alpha \frac{\partial^6 u}{\partial y^6}, \ t^\alpha \frac{\partial^6 u}{\partial y^4\, \partial x^2} \in H^\varepsilon(\overline{Q}). \tag{3.53}$$

Le lemme 3.4 se trouve démontré.

LEMME 3.5. *Hypothèses du théorème* 3.3. *La solution* v_1 *du problème* (3.43), (3.41), (3.42) *vérifie les relations*

$$t^\alpha \frac{\partial^2 v_1}{\partial t^2}, \ t^\alpha \frac{\partial^4 v_1}{\partial x^4}, \ t^\alpha \frac{\partial^4 v_1}{\partial y^4} \in H^\varepsilon(\overline{Q}).$$

DÉMONSTRATION. On pose $z = \partial v_1/\partial t$ et on reprend le schéma de démonstration (3.44) à (3.49) tout en tenant compte des résultats (3.50), (3.51). On obtient

$$t^\alpha \frac{\partial^2 v_1}{\partial t^2}, \ t^\alpha \frac{\partial^3 v_1}{\partial t\, \partial x^2}, \ t^\alpha \frac{\partial^3 v_1}{\partial t\, \partial y^2} \in H^\varepsilon(\overline{Q}). \tag{3.54}$$

Soit maintenant $z = \partial^2 v_1/\partial x^2$. Avec (3.52), (3.54), on a

$$t^\alpha \frac{\partial^4 v_1}{\partial x^4}, \ t^\alpha \frac{\partial^4 v_1}{\partial x^2\, \partial y^2} \in H^\varepsilon(\overline{Q}). \tag{3.55}$$

Ces raisonnements conduisent à la substitution $x = y$ près à la relation

$$t^\alpha \frac{\partial^4 v_1}{\partial y^4} \in H^\varepsilon(\overline{Q}), \tag{3.56}$$

qui achève la démonstration du lemme.

On continue la démonstration du théorème 3.3. Les relations (3.54) à (3.56) sont justes pour w_1 car $w_1 = v_1$. On choisit v_2, w_2 comme solution du problème

$$\frac{\partial v^2}{\partial t} = \frac{\partial^2 v_2}{\partial y^2} + \frac{\partial^2 w_2}{\partial x^2} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} \quad \text{dans} \quad Q, \tag{3.57}$$

$$w_2 = v_2 - \frac{\partial^2 u}{\partial y^2} \quad \text{dans} \quad Q, \tag{3.58}$$

$$\left.\begin{aligned} &v_2 = 0 \quad \text{sur } S, \\ &v_2(\mathbf{x}, 0) = 0, \qquad \mathbf{x} \in \Omega. \end{aligned}\right\} \tag{3.59}$$

On élimine w_2 de (3.57) moyennant la relation (3.58), il vient l'équation pour v_2:

$$\frac{\partial v_2}{\partial t} = \Delta v_2 - \frac{\partial^4 u}{\partial x^2\, \partial y^2} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} \quad \text{dans} \quad Q. \tag{3.60}$$

Le problème (3.59), (3.60) vérifie les hypothèses du théorème 3.1, où la constante $l = 1 + \lambda, \lambda \in (0, 1)$. Il existe donc une solution unique $v_2 \in H^{3+\lambda}(\overline{Q})$. On utilise la relation (3.58) et la régularité de la fonction u pour trouver $w_2 \in H^{3+\lambda}(\overline{Q})$. Leur indépendance par rapport à τ et h est évidente. On a pour la fonction v_2 le résultat suivant qui est un analogue de l'affirmation relative à v_1 du lemme précédent.

LEMME 3.6. *Hypothèses du théorème* 3.3. *On a*

$$t^\alpha \frac{\partial^2 v_2}{\partial t^2}, \ t^\alpha \frac{\partial^4 v_2}{\partial x^4}, \ t^\alpha \frac{\partial^4 v_2}{\partial y^4} \in H^\varepsilon(\overline{Q}).$$

On passe outre à la démonstration qui est en fait celle du lemme 3.5.

Connaissant u, v_i, w_i, on définit les fonctions discrètes

$$\eta^\tau = (u^h - u - h^2 v_1 - \tau v_2)\frac{1}{h^4 + \tau^2} \quad \text{sur} \quad Q_h^\tau \cup (\Gamma_{h,y} \times \omega_\tau),$$

$$\eta^* = (u^* - u - h^2 w_1 - \tau w_2)\frac{1}{h^4 + \tau^2} \quad \text{sur} \quad Q_h^\tau \cup (\Gamma_{h,x} \times \omega_\tau).$$

Il nous reste, pour en finir avec le théorème 3.3, à prouver les estimations (3.38). On substitue à u^τ et u^* de (3.12), (3.13) leurs développements (3.37):

$$\frac{u-u(x,y,t-\tau)}{\tau}+\frac{h^2}{\tau}(w_1-v_1(x,y,t-\tau))+$$
$$+(w_2-v_2(x,y,t-\tau))+(h^4+\tau^2)\frac{\eta^*-\eta^\tau(x,y,t-\tau)}{\tau}-$$
$$-L_1^h u-h^2L_1^h w_1-\tau L_1^h w_2-(h^4+\tau^2)L_1^h\eta^*=f, \quad (3.61)$$

$$\frac{h^2}{\tau}(v_1-w_1)+(v_2-w_2)+(h^4+\tau^2)\frac{\eta^\tau-\eta^*}{\tau}-$$
$$-L_2^h u-h^4L_2^h v_1-\tau L_2^h v_2-(h^4+\tau^2)L_2^h\eta^\tau=0. \quad (3.62)$$

On transforme (3.61) et (3.62) en utilisant la régularité des fonctions u, v_i, w_i. On fixe un certain $\alpha\in(\varepsilon,1)$. Avec les lemmes 3.4, 3.5 et 3.6, les fonctions ξ_i des développements

$$\frac{u-u(\mathbf{x},t-\tau)}{\tau}=\frac{\partial u}{\partial t}-\frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}+\tau^2\xi_1,$$
$$v_1(\mathbf{x},t-\tau)=v_1-\tau\frac{\partial v_1}{\partial t}-\tau^2\xi_2, \quad (3.63)$$
$$v_2(\mathbf{x},t-\tau)=v_2-\tau\frac{\partial v_2}{\partial t}+\tau^2\xi_3$$

vérifient les inégalités

$$|\xi_i(\mathbf{x},t)|\leqslant t^{-\alpha}c_2 \qquad \forall(\mathbf{x},t)\in Q_h^\tau, \qquad i=1,2,3.$$

On calcule $L_1^h u$. Si $\mathbf{x}\in\Omega_{h,x}^r$, on a par le lemme 1.3, § 7.1

$$L_1^h u=\frac{\partial^2 u}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}+h^4\xi_4, \quad (3.64)$$

où

$$|\xi_4|\leqslant c_3 t^{-\alpha} \qquad \forall(\mathbf{x},t)\in\Omega_{h,x}^r\times\omega_\tau.$$

Si $\mathbf{x}$ est irrégulier dans la direction x, alors la formule (2.29) du n° 4.2.3, ch. 4 entraîne

$$L_1^h u=\frac{\partial^2 u}{\partial x^2}+h^2\xi_5, \quad (3.65)$$

où

$$|\xi_5|\leqslant c^4 \qquad \forall(\mathbf{x},t)\in\Omega_{h,x}^{ir}\times\omega_\tau.$$

On réunit (3.64), (3.65) en une seule formule, à savoir

$$L_1^h u=\frac{\partial^2 u}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}+h^4\xi^9, \quad (3.66)$$

où

$$|\xi_6| \leqslant \begin{cases} c_3\, t^{-\alpha}, & (\mathbf{x},\, t) \in \Omega^{r}_{h,x} \times \omega_\tau, \\ \dfrac{c_4}{h^2}, & (\mathbf{x},\, t) \in \Omega^{ir}_{h,x} \times \omega_\tau. \end{cases} \tag{3.67}$$

On trouve de même

$$\begin{aligned} L_1^h w_1 &= \frac{\partial^2 w_1}{\partial x^2} + h^2 \xi_7, \\ L_1^h w_2 &= \frac{\partial^2 w_2}{\partial x^2} + h^2 \xi_8, \end{aligned} \tag{3.68}$$

où

$$|\xi_j| = \begin{cases} c_6\, t^{-\alpha}, & (\mathbf{x},\, t) \in \Omega^{r}_{h,x} \times \omega_\tau, \\ \dfrac{c_7}{h^2}, & (\mathbf{x},\, t) \in \Omega^{ir}_{h,x} \times \omega_\tau, \end{cases} \tag{3.69}$$

$$j = 7,\ 8.$$

Si l'on remplace x par y, alors

$$\begin{aligned} L_2^h u &= \frac{\partial^2 u}{\partial y^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial y^4} + h^4 \xi_9, \\ L_2^h v_1 &= \frac{\partial^2 v_1}{\partial y^2} + h^2 \xi_{10}, \\ L_2^h v_2 &= \frac{\partial^2 v_2}{\partial y^2} + h^2 \xi_{11}, \end{aligned} \tag{3.70}$$

où

$$|\xi_j| = \begin{cases} c_8\, t^{-\alpha}, & (\mathbf{x},\, t) \in \Omega^{r}_{h,y} \times \omega_\tau, \\ \dfrac{c_9}{h^2}, & (\mathbf{x},\, t) \in \Omega^{ir}_{h,y} \times \omega_\tau, \end{cases} \tag{3.71}$$

$$j = 9,\ 10,\ 11.$$

On transforme (3.61) moyennant les développements (3.63) à (3.68) et on met ensemble les termes contenant les mêmes puissances de τ et h:

$$\begin{aligned} \left(\frac{\partial u}{\partial t} + w_2 - v_2 - \frac{\partial^2 u}{\partial x^2}\right) &+ \tau\left(-\frac{1}{2}\frac{\partial^2 u}{\partial t^2} + \frac{\partial v_2}{\partial t} - \frac{\partial^2 w_2}{\partial x^2}\right) + \\ &+ \frac{h^2}{\tau}(w_1 - v_1) + h^2\left(\frac{\partial v_1}{\partial t} - \frac{1}{12}\frac{\partial^4 u}{\partial x^4} - \frac{\partial^2 w_1}{\partial x^2}\right) + \\ &+ \tau^2(\xi_1 - \xi_3) - h^4(\xi_6 + \xi_7) - \tau h^2(\xi_8 + \xi_2) + \\ &+ (h^4 + \tau^2)\left(\frac{\eta^* - \eta^*(x,\, y,\, t - \tau)}{\tau} - L_1^h \eta^*\right) = f, \end{aligned} \tag{3.72}$$

On simplifie à l'aide des relations précédentes. Les équations (3.1), (3.58) entraînent

$$\frac{\partial u}{\partial t} + w_2 - v_2 - \frac{\partial^2 u}{\partial x^2} = f.$$

Avec l'inégalité

$$\tau h^2 \leqslant \frac{1}{2}(\tau^2 + h^4). \tag{3.73}$$

on met la somme de toutes les expressions en ξ_j sous forme de $(h^4 + \tau^2)\,\xi_{12}$, où

$$|\xi_{12}| \leqslant \begin{cases} c_{10}\, t^{-\alpha}, & (\mathbf{x}, t) \in \Omega^r_{h,x} \times \omega_\tau, \\ c_{11}/h^2, & (\mathbf{x}, t) \in \Omega^{ir}_{h,x} \times \omega_\tau. \end{cases} \tag{3.74}$$

La relation (3.72) s'écrit donc

$$(h^4 + \tau)^2 \left(\frac{\eta^* - \eta^\tau(x, y, t - \tau)}{\tau} - L_1^h \eta^* \right) =$$
$$= -\tau \left(\frac{\partial v_2}{\partial t} - \frac{1}{2}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 w_2}{\partial x^2} \right) -$$
$$- h^2 \left(\frac{\partial v_1}{\partial t} - \frac{1}{12}\frac{\partial^4 u}{\partial x^4} - \frac{\partial^2 w_1}{\partial x^2} \right) - (h^4 + \tau^2)\,\xi_{12}. \tag{3.75}$$

On transforme maintenant l'égalité (3.62) à l'aide de (3.70) et on effectue un groupement analogue au cas précédent, il vient

$$\left(v_2 - w_2 - \frac{\partial^2 u}{\partial y^2} \right) + \frac{h^2}{\tau}(v_1 - w_1) + h^2 \left(-\frac{1}{12}\frac{\partial^4 u}{\partial y^4} - \frac{\partial^2 v_1}{\partial y^2} \right) -$$
$$- h^4(\xi_9 + \xi_{10}) - \tau \frac{\partial^2 v_2}{\partial y^2} - \tau h^2 \xi_{11} +$$
$$+ (h^4 + \tau^2) \left(\frac{\eta^\tau - \eta^*}{\tau} - L_2^h \eta^\tau \right) = 0. \tag{3.76}$$

On effectue des simplifications moyennant (3.40), (3.58) et on met la somme de tous les termes en ξ_j sous forme de $(h^4 + \tau^2)\,\xi_{13}$, où

$$|\xi_{13}| \leqslant \begin{cases} c_{12}\, t^{-\alpha}, & (\mathbf{x}, t) \in \Omega^r_{h,y} \times \omega_\tau, \\ c_{13}/h^2, & (\mathbf{x}, t) \in \Omega^{ir}_{h,y} \times \omega_\tau. \end{cases} \tag{3.77}$$

Ainsi, la relation (3.76) se ramène à

$$(h^4 + \tau^2) \left(\frac{\eta^\tau - \eta^*}{\tau} - L_2^h \eta^\tau \right) =$$
$$= \tau \frac{\partial^2 v_2}{\partial y^2} + h^2 \left(\frac{1}{12}\frac{\partial^4 u}{\partial y^4} + \frac{\partial^2 v_1}{\partial y^2} \right) - (h^4 + \tau^2)\,\xi_{13}. \tag{3.78}$$

On adjoint à cette équation les conditions initiales et aux limites qui découlent de la définition des fonctions η^τ et η^* et du fait que u^τ, u^*, u, v_i, w_i sont nulles aux nœuds correspondants:

$$\eta^* = 0 \quad \text{sur} \quad \Gamma_{h,x} \times \omega_\tau. \tag{3.79}$$

$$\eta^\tau = 0 \quad \text{sur} \quad \Gamma_{h,y} \times \omega_\tau. \tag{3.80}$$

$$\eta^\tau(\mathbf{x}, 0) = 0 \quad \forall\, \mathbf{x} \in \Omega_h. \tag{3.81}$$

Le problème (3.75), (3.78) à (3.81) ainsi obtenu vérifie l'estimation du théorème 3.2. Cette estimation permet d'établir l'unicité, les conditions de possibilité étant construites plus haut. Mais elle ne donne pas directement le résultat voulu car les seconds membres de (3.75) et (3.78) ne sont pas des quantités en h^4 et τ^2. On dégage explicitement les termes en τ et h^2 de η^τ et η^*, dont la contribution dans les seconds membres est en τ^2 et h^4. Ces composantes sont solution du problème

$$\frac{z^* - z^\tau(x, y, t - \tau)}{\tau} = -\tau\left(\frac{\partial v_2}{\partial t} - \frac{1}{2}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 w_2}{\partial x^2}\right) - $$
$$- h^2\left(\frac{\partial v_1}{\partial t} - \frac{1}{12}\frac{\partial^4 u}{\partial x^4} - \frac{\partial^2 w_1}{\partial x^2}\right) \quad \text{sur} \quad Q_h^\tau. \tag{3.82}$$

$$z^* = 0 \quad \text{sur} \quad \Gamma_{h,x} \times \omega_\tau. \tag{3.83}$$

$$\frac{z^\tau - z^*}{\tau} = \tau\frac{\partial^2 v_2}{\partial y^2} + h^2\left(\frac{1}{12}\frac{\partial^4 u}{\partial y^4} + \frac{\partial^2 v_1}{\partial y^2}\right) \quad \text{sur} \quad Q_h^\tau, \tag{3.84}$$

$$z^\tau = 0 \quad \text{sur} \quad \Gamma_{h,y} \times \omega_\tau \tag{3.85}$$

avec la condition initiale

$$z^\tau(\mathbf{x}, 0) = 0, \qquad \mathbf{x} \in \Omega_h. \tag{3.86}$$

On démontre que la solution du problème (3.82) à (3.86) s'écrit

$$z^\tau = 0 \quad \text{sur} \quad \overline{Q_h^\tau}. \tag{3.87}$$

$$z^* = \tau^2\left(-\frac{\partial v_2}{\partial t} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2 w_2}{\partial x^2}\right) +$$
$$+ \tau h^2\left(-\frac{\partial v_1}{\partial t} + \frac{1}{12}\frac{\partial^4 u}{\partial x^4} + \frac{\partial^2 w_1}{\partial x^2}\right) \quad \text{sur} \quad Q_h^\tau. \tag{3.88}$$

On a

$$z^\tau = 0 \quad \text{sur} \quad \Gamma_{h,x} \times \omega_\tau. \tag{3.89}$$

$$z^* = 0 \quad \text{sur} \quad \Gamma_{h,y} \times \omega_\tau. \tag{3.90}$$

Le report de ces fonctions dans (3.82) donne une identité. La condition initiale (3.86) est satisfaite car $z^\tau = 0$ partout sur $\overline{Q_h^\tau}$. Il reste à vérifier (3.84). On y substitue les fonctions (3.87), (3.88):

$$-\tau\left(-\frac{\partial v_2}{\partial t}+\frac{1}{2}\frac{\partial^2 u}{\partial t^2}+\frac{\partial^2 w_2}{\partial x^2}\right)-h^2\left(-\frac{\partial v_1}{\partial t}+\frac{1}{12}\frac{\partial^4 u}{\partial x^4}+\frac{\partial^2 w_1}{\partial x^2}\right)=$$

$$=\tau\frac{\partial^2 v_2}{\partial y^2}+h^2\left(\frac{1}{12}\frac{\partial^4 u}{\partial y^4}+\frac{\partial^2 v_1}{\partial y^2}\right).$$

C'est une identité du moment que la définition de v_i et w_i donne lieu à l'égalité

$$\frac{\partial v_2}{\partial t}-\frac{1}{2}\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 w_2}{\partial x^2}=\frac{\partial^2 v_2}{\partial y^2},$$

conséquence de (3.60), et

$$\frac{\partial v_1}{\partial t}-\frac{1}{12}\frac{\partial^4 u}{\partial x^4}-\frac{\partial^2 w_1}{\partial x^2}=\frac{1}{12}\frac{\partial^4 u}{\partial y^4}+\frac{\partial^2 v_1}{\partial y^2}$$

résultant de (3.43). Ainsi, les fonctions z^τ et z^* définies par les formules (3.87) à (3.90) sont solution du problème (3.82) à (3.86).

On effectue dans le problème (3.75), (3.78) à (3.81) la substitution

$$\zeta^\tau=\eta^\tau-\frac{1}{h^4+\tau^2}z^\tau,\qquad \zeta^*=\eta^*-\frac{1}{h^4+\tau^2}z^*,$$

auquel cas on obtient le système suivant pour ζ^τ:

$$(h^4+\tau^2)\left(\frac{\zeta^*-\zeta^\tau(x,y,t-\tau)}{\tau}-L_1^h\zeta^*\right)=$$

$$=-(h^4+\tau^2)\,\xi_{12}+L_1^h z^*\quad \text{sur}\quad Q_h^\tau. \tag{3.91}$$

$$\zeta^*=0\quad \text{sur}\quad \Gamma_{h,x}\times\omega_\tau. \tag{3.92}$$

$$(h^4+\tau^2)\left(\frac{\zeta^\tau-\zeta^*}{\tau}-L_2^h\zeta^\tau\right)=-(h^4+\tau^2)\,\xi_{13}\quad \text{sur}\quad Q_h^\tau, \tag{3.93}$$

$$\zeta^\tau=0\quad \text{sur}\quad \Gamma_{h,y}\times\omega_\tau, \tag{3.94}$$

$$\zeta^\tau(\mathbf{x},0)=0,\quad \mathbf{x}\in\Omega_h. \tag{3.95}$$

Il ne reste plus, pour justifier dans ce cas le théorème 3.2, qu'à évaluer $L_1^h z^*$ de (3.91).

Soit $\mathbf{x}\in\Omega'_{h,x}$ un point régulier dans la direction x. On a

$$L_1^h z^*=\dot z_{\bar x x}=\tau^2(g_1)_{\bar x x}+\tau h^2(g_2)_{\bar x x},$$

où

$$g_1=-\frac{\partial v_2}{\partial t}+\frac{1}{2}\frac{\partial^2 u}{\partial t^2}+\frac{\partial^2 w_2}{\partial x^2},$$

$$g_2=-\frac{\partial v_1}{\partial t}+\frac{1}{12}\frac{\partial^4 u}{\partial x^4}+\frac{\partial^2 w_1}{\partial x^2}.$$

Les résultats 3.4, 3.5, 3.6 et le lemme 1.3, § 7.1 permettent d'obtenir l'égalité

$$L_1^h z^* = \tau^2 \xi_{14} + \tau h^2 \xi_{15}. \tag{3.96}$$

avec

$$|\xi_j| \leqslant c_{14} t^{-\alpha} \qquad \forall (\mathbf{x}, t) \in \Omega_{h,x}^r \times \omega_\tau. \qquad j = 14, 15.$$

Soit $\mathbf{x} \in \Omega_{h,x}^{ir}$ un point irrégulier dans la direction x. On a

$$L_1^h z^* = \tau^2 L_1^h g_1 + \tau h^2 L_1^h g_2.$$

La définition (3.9) de L_1^h donne

$$L_1^h g_1(x, y, t) = -\frac{3-\delta}{\delta h^2} g_1(x, y, t) +$$
$$+ \frac{4 - 2\delta}{(1+\delta) h^2} g_1(x \pm h, y, t) - \frac{1-\delta}{(2+\delta) h^2} g_1(x \pm 2h, y, t).$$

La fonction g_1 étant continue sur $\overline{Q}$, on a l'inégalité $|g_1| \leqslant c_{15}$, d'où

$$|L_1^h g_1(\mathbf{x}, t)| \leqslant \frac{1}{h^2}\left(\frac{3-\delta}{\delta} + \frac{4-2\delta}{1+\delta} + \frac{|1-\delta|}{2+\delta}\right) c_{15}.$$

Comme $\delta \in (0, 3/2)$,

$$|L_1^h g_1(\mathbf{x}, t)| \leqslant \frac{1}{h^2}\left(\frac{3}{\delta} + 4 + \frac{1}{2}\right) c_{15} \leqslant \frac{10\, c_{15}}{h^2 \delta}.$$

Il en résulte

$$\tau^2 L_1^h g_1 = \frac{\tau^2}{h^2} \xi_{16},$$

avec

$$|\xi_{16}| \leqslant \frac{c_{16}}{\delta} \qquad \forall (\mathbf{x}, t) \in \Omega_{h,x}^{ir} \times \omega_\tau, \qquad \delta = \delta(\mathbf{x}).$$

On procède de même en ce qui concerne $\tau h^2 L_1^h g_2$. On a

$$L_1^h z^* = (\tau^2 + \tau h^2) \frac{\xi_{17}}{h^2}, \tag{3.97}$$

où

$$|\xi_{17}| \leqslant c_{17}/\delta \qquad \forall (\mathbf{x}, t) \in \Omega_{h,x}^{ir} \times \omega_\tau.$$

On réunit (3.96) et (3.97):

$$L_1^h z^* = (\tau^2 + h^4) \xi_{18}, \tag{3.98}$$

avec

$$|\xi_{18}| \leqslant \begin{cases} 2\, c_{14}\, t^{-\alpha}, & (\mathbf{x}, t) \in \Omega_{h,x}^r \times \omega_\tau, \\ \dfrac{2 c_{17}}{h^2 \delta}, & (\mathbf{x}, t) \in \Omega_{h,x}^r \times \omega_\tau. \end{cases} \tag{3.99}$$

Aussi l'équation (3.91) s'écrit

$$(\tau^2 + h^4)\left(\frac{\zeta^t - \zeta^\tau(x, t-\tau)}{\tau} - L_1^h \zeta^*\right) = (\tau^2 + h^4)(\xi_{13} - \xi_{12}) \quad \text{sur} \quad Q_h^\tau. \quad (3.100)$$

On divise (3.93) et (3.100) membre à membre par $h^4 + \tau^2$ et on est conduit à un problème dont la solution vérifie l'estimation du théorème 3.2. On a par suite de cette estimation

$$\max_{Q_h^\tau} |\zeta^*| \leqslant \sum_{t\in\omega_\tau} \tau \left(\max_{\Omega_{h,x}^r} |\xi_{12}| + \max_{\Omega_{h,x}^r} |\xi_{18}| + \max_{\Omega_{h,x}^r} |\xi_{13}|\right) + $$
$$+ 2h^2 \max\left\{ \max_{\Omega_{h,x}^{ir}\times\omega_\tau} |\delta\xi_{12}| + \max_{\Omega_{h,x}^{ir}\times\omega_\tau} |\delta\xi_{13}| + \max_{\Omega_{h,y}^{ir}\times\omega_\tau} |\rho\xi_{13}|\right\},$$

d'où l'on déduit moyennant (3.74), (3.77) et (3.99):

$$\max_{Q_h^\tau} |\zeta^*| \leqslant \sum_{t\in\omega_\tau} \tau t^{-\alpha} (c_{10} + c_{12} + 2c_{14}) + $$
$$+ 2\left(\max_{\Omega_{h,x}^{ir}\times\omega_\tau} c_{11}\delta + 2c_{17} + \max_{\Omega_{h,y}^{ir}\times\omega_\tau} c_{13}\rho\right). \quad (3.101)$$

Comme $\alpha < 1$, on a

$$\sum_{t\in\omega_\tau} \tau t^{-\alpha} \leqslant \int_0^T t^{-\alpha}\, dt = \frac{T^{1-\alpha}}{1-\alpha}.$$

Cette inégalité plus les estimations $\delta(x) \leqslant 3/2$, $\rho(x) \leqslant 3/2$ permettent de simplifier (3.101):

$$\max_{Q_h^\tau} |\zeta^*| \leqslant c_{18}, \quad (3.102)$$

où

$$c_{18} = \frac{T^{1-\alpha}}{1-\alpha}(c_{10} + c_{12} + 2c_{14}) + 3c_{11} + 3c_{13} + 2c_{17}.$$

On trouve de même pour ζ^τ:

$$\max_{\overline{Q}_h^\tau} |\zeta^\tau| \leqslant c_{18}.$$

Etant données les relations

$$\eta^\tau = \zeta^\tau, \qquad \eta^* = \zeta^* + \frac{z^t}{\tau^2 + h^4}.$$

l'égalité (3.88), la propriété de tous les termes de (3.88) d'être continus sur $\overline{Q}$ et l'inégalité (3.73), on trouve les estimations (3.38), i.e. la dernière affirmation du théorème 3.3.

On va décrire une méthode de raffinement basée sur le développement prouvé. Supposons qu'on est dans les conditions du théorème 3.3. On choisit $M \geqslant 2$ et $N \geqslant 2$ entiers et on construit le réseau de discrétisation $\overline{Q}_h^\tau$ de pas de temps τ et de pas d'espace h. On cherche la solution u^τ du problème (3.12) à (3.14). On construit le réseau $\overline{Q}_{h/2}^{\tau/4}$ de pas $\tau/4$ et $h/2$, on cherche la solution du problème et on la note $u^{\tau/4}$. Il y a deux solutions approchées associées aux nœuds de $\overline{Q}_h^\tau$. On forme la combinaison linéaire

$$U^H(\mathbf{x}, t) = \frac{4}{3} u^{\tau/4}(\mathbf{x}, t) - \frac{1}{3} u^\tau(\mathbf{x}, t), \qquad (\mathbf{x}, t) \in \overline{Q}_h^\tau, \qquad (3.103)$$

et on démontre que la solution améliorée U^H approche en norme uniforme la solution exacte u avec une précision en $\tau^2 + h^4$.

THÉORÈME 3.7. *Hypothèses du théorème* 3.3. *La solution corrigée* (3.103) *admet l'estimation*

$$\max_{\overline{Q}_h^\tau} |U^H - u| \leqslant \frac{5}{12} c_1 (\tau^2 + h^4). \qquad (3.104)$$

DÉMONSTRATION. On a en chaque nœud de $\overline{Q}_h^\tau$

$$u^\tau = u + h^2 v_1 + \tau v_2 + (\tau^2 + h^4)\eta^\tau,$$

$$u^{\tau/4} = u + \frac{h^2}{4} v_1 + \frac{\tau}{4} v_2 + \frac{1}{16}(\tau^2 + h^4)\eta^{\tau/4}.$$

Les fonctions v_1, v_2 sont indépendantes de τ et h, si bien que

$$U^H = u + (\tau^2 + h^4)\frac{1}{3}\left(\frac{1}{4}\eta^{\tau/4} - \eta^\tau\right).$$

Etant donnée l'estimation (3.38), on obtient

$$|U^H(\mathbf{x}, t) - u(\mathbf{x}, t)| \leqslant \frac{1}{3}(\tau^2 + h^4)\frac{5}{4} c_1, \qquad (\mathbf{x}, t) \in \overline{Q}_h^\tau.$$

c.q.f.d.

Revenons au schéma localement de dimension un (voir [43]). L'approximation de $\partial^2/\partial x^2$ et $\partial^2/\partial y^2$ au voisinage de la frontière est plus simple que celle effectuée avec le schéma décrit (elle correspond au procédé pour $n = 1$ du n° 4.2.3). Les idées de base des démonstrations restent les mêmes, et les calculs se simplifient sensiblement. Nous nous abstiendrons de démontrer le résultat fondamental énoncé sous forme de

THÉORÈME 3.8. *Hypothèses du théorème* 3.3. *La solution corrigée* (3.103) *formée de solutions du schéma localement de dimension un vérifie l'estimation*

$$\max_{\bar{\varphi}_h^{\tau}} | U^H - u | \leqslant c_{19} (\tau^2 + h^3). \tag{3.105}$$

5.4. Equation du mouvement

Dans ce paragraphe, nous nous occuperons du problème de Cauchy pour l'équation du mouvement

$$\begin{gathered} \frac{\partial \Phi}{\partial t} + \sum_{\alpha=1}^{p} v_\alpha (t, \mathbf{x}) \frac{\partial \Phi}{\partial x_\alpha} = f(t, \mathbf{x}), \qquad (t, \mathbf{x}) \in T \times \mathbf{R}^p, \\ \Phi(0, \mathbf{x}) = g(\mathbf{x}), \qquad \mathbf{x} = (x_1, \ldots, x_p) \in \mathbf{R}^p, \end{gathered} \tag{4.1}$$

où $T = [0, T_0]$. On suppose que les coefficients v_α vérifient l'équation de continuité

$$\sum_{\alpha=1}^{p} \frac{\partial v_\alpha}{\partial x_\alpha} = 0 \quad \text{dans} \quad T \times \mathbf{R}^p. \tag{4.2}$$

L'équation du mouvement présente une propriété de valeur qui permet de passer du problème dans un demi-espace à un autre dans un domaine borné. Le fait est que la relation entre la solution, d'une part, et les données initiales et le second membre, de l'autre, est déterminée par les caractéristiques de l'équation différentielle (voir [78]), si bien qu'on trouve la solution dans le domaine concerné sans la chercher dans l'espace tout entier. En effet, soit $\varkappa_\sigma^{\mathbf{x}_0}(\mathbf{x}) \in C^\infty(\mathbf{R}^p)$ une fonction de troncature égale à 1 dans la boule $B_\sigma(\mathbf{x}_0)$ de centre $\mathbf{x}_0 \in \mathbf{R}^p$ et de rayon $\sigma > 0$ et nulle en dehors de la boule $B_{2\sigma}(\mathbf{x}_0)$. Si les vitesses v_α sont finies partout dans $\mathbf{R}^p$, on trouve à l'instant t pour tout $\Omega \subset \mathbf{R}^p$ un point $\mathbf{x}_0 \in \Omega$ et σ suffisamment grand tels que la solution du problème (4.1), (4.2) coïncide sur Ω avec celle de

$$\begin{gathered} \frac{\partial \Phi_1}{\partial t} + \sum_{\alpha=1}^{p} v_\alpha \frac{\partial \Phi_1}{\partial x_\alpha} = f \varkappa_\sigma^{\mathbf{x}_0} \quad \text{dans} \quad T \times \mathbf{R}^p, \\ \Phi_1 (0, \mathbf{x}) = g(\mathbf{x}) \varkappa_\sigma^{\mathbf{x}_0}(\mathbf{x}), \qquad \mathbf{x} \in \mathbf{R}^p. \end{gathered}$$

La classe des fonctions continues dans Q ayant dans Q des dérivées continues jusqu'à l'ordre k sera désignée par la notation usuelle $C^k(Q)$, k entier. On écrira $\varphi \in \tilde{C}^k(Q)$ pour une fonction $\varphi \in C^k(Q)$ à support borné. Une fonction de classe $C^k(Q)$ est bornée, ainsi que toutes ses dérivées partielles jusqu'à l'ordre k inclus.

On se place dans le cas où les données initiales et le second membre possèdent des supports locaux:

$$v_\alpha(t, \mathbf{x}) \in C^{2r}(T \times \mathbf{R}^p), \qquad f(t, \mathbf{x}) \in \tilde{C}^{2r}(T \times \mathbf{R}^p),$$
$$g(\mathbf{x}) \in \tilde{C}^{2r}(\mathbf{R}^p), \tag{4.3}$$

r étant un entier naturel. Le problème (4.1), (4.2) admet alors une solution unique (voir [78]), et la solution Φ est dans $\tilde{C}^{2r}(T \times \mathbf{R}^p)$.

On suppose maintenant que la fonction Φ a son support dans le parallélépipède $Q = \{(t, x_1, \ldots, x_p),\ 0 \leqslant t \leqslant T_0,\quad 0 < x_\alpha < 1,\ \alpha = 1, \ldots, p\}$. On y arrive toujours par une transformation linéaire des variables d'espace. Dans ce cas, Φ est nulle sur la surface latérale S du parallélépipède:

$$x_\alpha = 0, \qquad x_\alpha = 1, \qquad \alpha = 1, \ldots, p.$$

On discrétise en les variables d'espace et de temps à l'aide des réseaux uniformes respectifs

$$\bar{\Omega}_h = \{\mathbf{x} = (x_1, \ldots, x_p) : x_\alpha = i_\alpha h,\ i_\alpha = 0, 1, \ldots, N\},$$
$$\Omega_h = \{\mathbf{x} \in \bar{\Omega}_h : i_\alpha = 1, \ldots, N-1 ;\ \alpha = 1, 2, \ldots, p\} \tag{4.4}$$

et

$$\bar{\omega}_\tau = \{t_j = \tau j,\ j = 0, \ldots, M\},$$
$$\omega_\tau = \{t \in \bar{\omega}_\tau ;\ t \neq 0\},$$

avec $\tau = T_0/M$, $h = 1/N$. On suppose qu'il y a entre les pas la relation

$$\tau = c_0 h, \tag{4.5}$$

c_0 étant une constante indépendante de τ et h.

On note $\overline{Q_h^\tau}$ le produit cartésien de réseaux $\bar{\omega}_\tau \times \bar{\Omega}_h$ et Q_h^τ le produit cartésien de réseaux $\omega_\tau \times \Omega_h$. On pose $S_h^\tau = \overline{Q_h^\tau} \cap S$.

On opère selon un schéma décomposé implicite et on obtient le problème

$$(I + \tau\Lambda_1) \ldots (I + \tau\Lambda_p)\, \Phi_\tau(t, \mathbf{x}) - \Phi_\tau(t - \tau, \mathbf{x}) = \tau f(t, \mathbf{x}),$$
$$(t, \mathbf{x}) \in Q_h^\tau, \tag{4.6}$$

$$\Phi_\tau(0, \mathbf{x}) = g(\mathbf{x}), \quad \mathbf{x} \in \Omega_h, \quad \Phi_\tau(t, \mathbf{x}) = 0, \quad (t, \mathbf{x}) \in S_h^\tau, \tag{4.7}$$

approchant le problème (4.1) à (4.3). Ici I est l'opérateur unité,

$$\Lambda_\alpha \varphi(t, \mathbf{x}) = (1/4)\, h\, \{(v_\alpha(t, x_1, \ldots, x_\alpha + h, \ldots, x_p) +$$
$$+ v_\alpha(t, x_1, \ldots, x_p))\ \varphi(t, x_1, \ldots, x_\alpha + h, \ldots, x_p) -$$
$$- (v_\alpha(t, x_1, \ldots, x_\alpha - h, \ldots, x_p) + v_\alpha(t, x_1, \ldots, x_p)) \times$$
$$\times\ \varphi(t, x_1, \ldots, x_\alpha - h, \ldots, x_p)\}, \ (t, \mathbf{x}) \in \omega_\tau \times \Omega_h.$$

On introduit pour les fonctions discrètes de domaine de définition Ω_h le produit scalaire

$$(a, b) = \sum_{x \in \Omega_h} a(\mathbf{x})\, b(\mathbf{x})\, h^p$$

et la norme $\| a \| = (a, a)^{1/2}$.

LEMME 4.1. *Toute fonction discrète φ définie sur $\bar{\Omega}_h$ et nulle sur $\bar{\Omega}_h \setminus \Omega_h$ vérifie la relation*

$$(\Lambda_\alpha \varphi, \varphi) = 0, \quad \alpha = 1, \ldots, p;\ t \in \omega_\tau. \tag{4.8}$$

DÉMONSTRATION. L'égalité $(\Lambda_\alpha \varphi, \varphi) = -(\varphi, \Lambda_\alpha \varphi)$ équivalente à (4.8) est un analogue discret de la formule d'intégration par parties, et on la vérifie de façon immédiate.

La propriété établie entraîne de suite le

LEMME 4.2. *Toute fonction discrète φ définie sur $\bar{\Omega}_h$ et nulle sur $\bar{\Omega}_h \setminus \Omega_h$ vérifie pour tout τ, $t \in T$, les inégalités*

$$\| \varphi \| \leqslant \| (I + \tau \Lambda_\alpha) \varphi \|, \quad \alpha = 1, 2, \ldots, p.$$

La démonstration en est donnée (à un changement de notations près) dans [30].

Ci-dessous deux résultats auxiliaires sur un développement particulier de l'erreur d'approximation.

LEMME 4.3. *On suppose que $\varphi \in \tilde{C}^q(T \times \mathbf{R}^p)$, $q \geqslant 2$, et qu'elle a son support à l'intérieur du parallélépipède Q. On a*

$$\Lambda_\alpha \varphi = \frac{\partial \varphi}{\partial x_\alpha} v_\alpha + \frac{1}{2} \varphi \frac{\partial v_\alpha}{\partial x_\alpha} + \sum_{i=1}^{q-2} \tau^i \varphi_i + \tau^{q-1} \varphi_{q-1,\tau} \ \textit{sur}\ Q_h^\tau.$$

Ici $\varphi_i \in \tilde{C}^{q-i-1}(T \times \mathbf{R}^p)$, $i = 1, \ldots, q-2$, ne dépend pas de τ, le support de la fonction φ_i étant intérieur à Q. La fonction discrète $\varphi_{q-1,\tau}$ est bornée :

$$\max_{Q_h^\tau} | \varphi_{q-1,\tau} | \leqslant c_2.$$

Si ψ est une fonction discrète quelconque définie sur $\bar{\Omega}_h$, alors

$$\max_{\Omega_h} | \Lambda_\alpha \psi | \leqslant \frac{c_3}{\tau} \max_{\bar{\Omega}_h} | \psi |.$$

DÉMONSTRATION. La dernière affirmation découle de la forme de l'opérateur Λ_α, et la constante c_3 est égale à

$$\frac{c_0}{2} \max_{\bar{Q}} | v_\alpha |.$$

On démontre le développement suivant les puissances de τ en mettant Λ_α sous la forme

$$\Lambda_\alpha \varphi(t, \mathbf{x}) = (1/4)\, h \{ v_\alpha \varphi|_{(t, x_1, \ldots, x_\alpha + h, \ldots, x_p)} - v_\alpha \varphi|_{(t, x_1, \ldots, x_\alpha - h, \ldots, x_p)} +$$

$$+ v_\alpha(t, \mathbf{x}) \{ \varphi(t, x_1, \ldots, x_\alpha + h, \ldots, x_p) - \varphi(t, x_1, \ldots, x_\alpha - h, \ldots, x_p)\}\}$$

et en appliquant deux fois le lemme 1.1, § 7.1, il vient le développement voulu, avec

$$\varphi_i = \begin{cases} 0 & \text{si } i \text{ est impair,} \\ \dfrac{c_0^{-i}}{(2i+1)!} \left\{ \dfrac{\partial^{i+1}(v_\alpha \varphi)}{\partial x_\alpha^{i+1}} + v_\alpha \dfrac{\partial^{i+1} \varphi}{\partial x_\alpha^{i+1}} \right\} & \text{si } i \text{ est pair, dans } T \times \mathbf{R}^p, \end{cases}$$

$$|\varphi_{q-1,\tau}| \leqslant \frac{c_0^{-q+1}}{2q!} \left(\max_{T \times \mathbf{R}^p} \left| \frac{\partial^q (v_\alpha \varphi)}{\partial x_\alpha^q} \right| + \max_{T \times \mathbf{R}^p} \left| v_\alpha \frac{\partial^q \varphi}{\partial x_\alpha^q} \right| \right) \text{ dans } Q_h^\tau .$$

LEMME 4.4. *Si l'on est dans les conditions de régularité* (4.3), *alors la solution du problème* (4.1), (4.2) *satisfait à la relation*

$$(I + \tau\Lambda_1) \ldots (I + \tau\Lambda_p)\, \Phi(t, \mathbf{x}) - \Phi(t - \tau, \mathbf{x}) =$$

$$= \tau \left\{ f + \sum_{i=1}^{r-1} \tau^i f_i + \tau^r f_{r,\tau} \right\} \Bigg|_{(t, x)}, \quad (t, \mathbf{x}) \in Q_h^\tau.$$

Ici les fonctions $f_i \in \tilde{C}^{r-i}(T \times \mathbf{R}^p)$, $i = 1, \ldots, r-1$, *sont indépendantes de* τ *et ont leur support à l'intérieur de* Q. *La fonction* $f_{r,\tau}$ *est bornée:*

$$\max_{Q_h^\tau} |f_{r,\tau}| \leqslant c_4.$$

DÉMONSTRATION. Une application successive du lemme 4.3 donne

$$(I + \tau\Lambda_1) \ldots (I + \tau\Lambda_p)\, \Phi = \Phi + \tau \sum_{\alpha=1}^{p} \left(v_\alpha \frac{\partial \Phi}{\partial x_\alpha} + \frac{1}{q} \Phi \frac{\partial v_\alpha}{\partial x_\alpha} \right) +$$

$$+ \sum_{i=2}^{r} \tau^i F_i + \tau^{r+1} F_{r+1,\tau} \text{ dans } Q_h^\tau .$$

Ici les fonctions $F_i \in \tilde{C}^{r+1-i}(T \times \mathbf{R}^p)$, $i = 2, \ldots, r$, sont indépendantes de τ et ont leur support à l'intérieur de Q, et la fonction discrète $F_{r+1,\tau}$ est bornée:

$$\max_{Q_h^\tau} |F_{r+1,\tau}| \leqslant c_5.$$

Etant donnée la condition (4.2), le coefficient de τ vaut

$$\sum_{\alpha=2}^{p} v_\alpha \frac{\partial \Phi}{\partial x_\alpha}.$$

On applique la formule de Taylor à $\Phi(t-\tau, \mathbf{x})$, il vient en définitive

$$(I + \tau \Lambda_1) \ldots (I + \tau \Lambda_p) \Phi(t, \mathbf{x}) - \Phi(t - \tau, \mathbf{x}) =$$

$$= \tau \left\{ \frac{\partial \Phi}{\partial t} + \sum_{\alpha=1}^{p} \frac{\partial \Phi}{\partial x_\alpha} v_\alpha + \sum_{i=2}^{r} \tau^{i-1} \overline{F}_i + \tau^r \overline{F}_{r+1,\tau} \right\} \Bigg|_{(t,x)}, \quad (t, \mathbf{x}) \in Q_h^\tau.$$

Ici on a pour $i = 2, \ldots, r$

$$\overline{F}_i = F_i + \frac{(-1)^{i-1}}{i!} \frac{\partial^i \Phi}{\partial t^i} \quad \text{sur} \quad T \times \mathbf{R}^p,$$

$$|\overline{F}_{r+1,\tau}(t, \mathbf{x})| \leqslant |F_{r+1,\tau}(t, \mathbf{x})| + \frac{1}{(r+1)!} \max_{T \times \mathbf{R}^p} \left| \frac{\partial^{r+1} \Phi}{\partial t^{r+1}} \right|,$$

$$(t, \mathbf{x}) \in Q_h^\tau.$$

On a l'affirmation du lemme du moment que τ a son coefficient égal à $f(t, \mathbf{x})$.

LEMME 4.5. *On a pour le schéma aux différences* (4.6), (4.7) *l'estimation à priori*

$$\| \Phi_\tau(t, \mathbf{x}) \| \leqslant \| g(\mathbf{x}) \| + T_0 \max_{t \in \omega_\tau} \| f(t, \mathbf{x}) \|, \quad t \in \omega_\tau.$$

La démonstration s'inspire du lemme 4.2, et on procède par récurrence sur $t_j \in \omega_\tau$ (voir par exemple [112]).

THÉORÈME 4.6. *On suppose que le problème différentiel* (4.1), (4.2.) *vérifie les conditions* (4.3) *et que les pas du problème aux différences* (4.6), (4.7) *satisfont à l'égalité* (4.5). *Les solutions de ces problèmes sont telles qu'on ait*

$$\Phi_\tau(t, \mathbf{x}) = \Phi(t, \mathbf{x}) + \sum_{i=1}^{r-1} \tau^i \Phi_i(t, \mathbf{x}) + \tau^r \Phi_{r,\tau}(t, \mathbf{x}), \qquad (t, \mathbf{x}) \in Q_h^\tau.$$

Ici les fonctions $\Phi_i \in \tilde{C}^{2r-2i}(T \times \mathbf{R}^p)$, $i = 1, \ldots, r-1$, *sont indépendantes de* τ (*et* h) *et les supports sont concentrés dans* Q. *La fonction discrète* $\Phi_{r,\tau}$ *est bornée*:

$$\max_{t \in \bar{\omega}_\tau} \| \Phi_{r,\tau} \| \leqslant c_6. \tag{4.9}$$

Démonstration. On a, vu le lemme 4.2 et les formules (4.6), (4.7) pour la solution aux différences,

$$(I + \tau \Lambda_1) \dots (I + \tau \Lambda_p) (\Phi_\tau - \Phi) |_{(t,x)} - (\Phi_\tau - \Phi) |_{(t-\tau,x)} =$$
$$= -\tau \left(\sum_{i=1}^{r-1} \tau^i f_i + \tau^r f_{r,\tau} \right) |_{(t,x)}, \qquad (t, \mathbf{x}) \in Q_h^\tau, \quad (4.10)$$
$$(\Phi_\tau - \Phi) |_{(0,x)} = 0, \qquad \mathbf{x} \in \Omega_h,$$
$$(\Phi_\tau - \Phi) |_{(t,x)} = 0, \qquad (t, \mathbf{x}) \in S_h^\tau.$$

On prend $\Phi_1(t, \mathbf{x})$ pour solution du problème

$$\frac{\partial \Phi_1}{\partial t} + \sum_{\alpha=1}^{p} v_\alpha \frac{\partial \Phi_1}{\partial x_\alpha} = -f_1(t, \mathbf{x}) \text{ sur } T \times \mathbf{R}^p,$$
$$\Phi_1 (0, \mathbf{x}) = 0, \ \mathbf{x} \in \mathbf{R}^p.$$

Comme $f_1 \in \tilde{C}^{2r-2} (T \times \mathbf{R}^p)$ et ne dépend pas de τ, il en est de même de $\Phi_1 \in \tilde{C}^{2r-2} (T \times \mathbf{R}^p)$. Il y a plus. La propriété de f_1 d'être non nulle uniquement le long des caractéristiques intérieures à Q est également celle de Φ_1. Aussi la dernière fonction a son support concentré à l'intérieur de Q. Cela autorise à utiliser le lemme 4.4, d'où

$$(I + \tau \Lambda_1) \dots (I + \tau \Lambda_p) \Phi_1 (t, \mathbf{x}) - \Phi_1(t - \tau, \mathbf{x}) =$$
$$= -\tau \left(f_1 + \sum_{i=1}^{r-2} \tau^i \bar{f}_i + \tau^{r-1} \bar{f}_{r,\tau} \right) \Bigg|_{(t,x)}, \qquad (t, \mathbf{x}) \in Q_h^\tau, \quad (4.11)$$
$$\Phi_1 (0, \mathbf{x}) = 0, \qquad \mathbf{x} \in \Omega_h,$$
$$\Phi_1 (t, \mathbf{x}) = 0, \qquad (t, \mathbf{x}) \in S_h^\tau.$$

On retranche les relations (4.11) multipliées par τ des égalités (4.10) correspondantes :

$$(I + \tau \Lambda_1) \dots (I + \tau \Lambda_p) (\Phi_\tau - \Phi - \tau \Phi_1) |_{(t,x)} -$$
$$- (\Phi_\tau - \Phi - \tau \Phi_1) |_{(t-\tau,x)} = -\tau \left(\sum_{i=2}^{r-1} \tau^i \tilde{f}_i + \tau^r \tilde{f}_{r,\tau} \right) \Bigg|_{(t,x)},$$
$$(t, \mathbf{x}) \in Q_h^\tau, \qquad (4.12)$$
$$(\Phi_\tau - \Phi - \tau \Phi_1) |_{(0,x)} = 0, \qquad \mathbf{x} \in \Omega_h,$$
$$(\Phi - \Phi - \tau \Phi_1) |_{(t,x)} = 0, \qquad (t, \mathbf{x}) \in S_h^\tau.$$

On choisit $r - 2$ Φ_i restantes de façon à obtenir après le $(r - 1)$-ième pas

$$(I + \tau\Lambda_1)\dots(I + \tau\Lambda_p)\left(\Phi_\tau - \Phi - \sum_{i=1}^{r-1} \tau^i \Phi_i\right)\Bigg|_{(t,x)} -$$

$$-\left(\Phi_\tau - \Phi - \sum_{i=1}^{r-1} \tau^i \Phi_i\right)\Bigg|_{(t-\tau,x)} = -\tau^{r+1}\hat{f}_{r,\tau}(t, \mathbf{x}).$$

$$(t, \mathbf{x}) \in Q_h^\tau,$$

$$\left(\Phi_\tau - \Phi - \sum_{i=1}^{r-1} \tau^i \Phi_i\right)\Bigg|_{(0,x)} = 0, \qquad \mathbf{x} \in \Omega_h.$$

$$\left(\Phi_\tau - \Phi - \sum_{i=1}^{r-1} \tau^i \Phi_i\right)\Bigg|_{(t,x)} = 0, \qquad (t, \mathbf{x}) \in S_h^\tau,$$

et

$$\max_{Q_h^\tau} |\hat{f}_{r,\tau}| \leqslant c_\tau.$$

d'où

$$\max_{t\in\omega_\tau} \|\hat{f}_{r,\tau}\| \leqslant c_\tau.$$

L'affirmation du théorème vient en appliquant à la fonction discrète

$$\tau\,\Phi_{r,\tau} = \Phi_\tau - \Phi - \sum_{i=1}^{r-1} \tau^i \Phi_i$$

l'estimation à priori du lemme 4.5.

Le théorème démontré aidant, on va décrire une méthode de raffinement de la solution discrète.

On suppose qu'on est, pour le problème (4.1), (4.2), dans les conditions (4.3). On fixe une constante c_0 et on construit pour $M_1 < M_2 < \dots < M_r$ entiers les réseaux de discrétisation en temps $\omega_{\tau_1}, \omega_{\tau_2}, \dots, \omega_{\tau_r}$ et les réseaux de discrétisation en espace Ω_{h_1}, $\Omega_{h_2}, \dots, \Omega_{h_r}$. Il y a entre les pas la relation

$$\tau_i = c_0 h_i \quad i = 1, \dots, r.$$

On résout pour chaque Q_h^τ le problème (4.6), (4.7), ce qui donne r solutions $\Phi_{\tau_1}, \dots, \Phi_{\tau_r}$. On se place dans le cas $Q_{h_1}^{\tau_1} \subset Q_{h_i}^{\tau_i}$ $i = 1,\dots$

.... r. On prend en chaque point de $Q_{h_1}^{\tau_1}$ une combinaison linéaire des Φ_{τ_i}:

$$\overline{\Phi}(t,\ \mathbf{x}) = \sum_{i=1}^{r} \gamma_i\, \Phi_{\tau_i}(t,\ \mathbf{x}), \qquad (t,\ \mathbf{x}) \in Q_{h_1}^{\tau_1}. \tag{4.13}$$

THÉORÈME 4.7. *On suppose que les données et le second membre du problème* (4.1), (4.2) *vérifient les conditions* (4.3). *Si M_i est tel qu'on ait l'estimation*

$$c_9 \geqslant \frac{M_{i+1}}{M_i} \geqslant 1 + c_8,\ i = 1,\ \ldots,\ r-1, \tag{4.14}$$

et si γ_i vérifient le système

$$\sum_{i=1}^{r} \gamma_i = 1,$$
$$\sum_{i=1}^{r} \gamma_i \frac{1}{M_i^l} = 0, \qquad l = 1,\ \ldots,\ r-1, \tag{4.15}$$

alors la solution obtenue par la formule (4.13) *admet la majoration*

$$|\,\Phi(t,\mathbf{x}) - \overline{\Phi}(t,\mathbf{x})\,| \leqslant \left(\sum_{i=1}^{r} |\,\gamma_i\,|\,\tau_i^r\right) b\,(t,\ \mathbf{x}),$$

$$(t,\ \mathbf{x}) \in Q_{h_1}^{\tau_1},$$

les nombres γ_1 étant bornés et la fonction discrète b étant bornée en moyenne quadratique:

$$\max_{t \in \bar{\omega}_{\tau_1}} \Big(\sum_{\mathbf{x} \in \Omega_{h_1}} |\,b\,(t,\ \mathbf{x})\,|^2\, h_1^p\Big)^{1/2} \leqslant c_{10}$$

où la constante c_{10} est indéndante de τ et h.

DÉMONSTRATION. Le fait de choisir γ_i à partir du système (4.15) et le théorème 4.6 garantissent l'inégalité

$$|\Phi(t,\ \mathbf{x}) - \overline{\Phi}(t,\ \mathbf{x})\,| \leqslant \sum_{i=1}^{r} |\,\gamma_i\,|\,\tau_i^r\,|\,\Phi_{r,\tau_i}\ (t,\ \mathbf{x})\,|.$$

$$(t,\ \mathbf{x}) \in Q_{h_1}^{\tau_1},$$

avec $|\,\gamma_i\,|$ bornés aux termes du lemme 2.3, § 7.2 par suite de la condition (4.14). La même condition fournit l'estimation

$$h_1/h_r \leqslant c_9^{r-1}.$$

Aussi

$$\left(\sum_{x\in\Omega_{h_i}} |\Phi_{r,\tau_i}(t,\ x)|^2 h_i^p\right)^{1/2} \geqslant \left(\sum_{x\in\Omega_{h_1}} |\Phi_{r,\tau_i}(t,\ x)|^2 h_i^p\right)^{1/2} \geqslant$$

$$\geqslant c_9^{(1-r)\,p/2}\left(\sum_{x\in\Omega_{h_1}} |\Phi_{r,\tau_i}(t,\ x)|^2 h_1^p\right)^{1/2}.$$

Avec l'inégalité (4.9), on est conduit à l'estimation

$$\left(\sum_{x\in\Omega_{h_1}} |\Phi_{r,\tau_i}(t,\ x)|^2 h_1^p\right)^{1/2} \leqslant c_6 c_9^{(r-1)p/2}.$$

Le théorème se trouve complètement démontré si l'on l'utilise pour évaluer la différence $\Phi - \overline{\Phi}$.

Remarque. On note que le théorème 4.6 entraîne un développement analogue (pour $r \geqslant p/2 + 1$) avec reste uniformément borné

$$\Phi_\tau(t,\ x) = \Phi(t,\ x) + \sum_{i=1}^{s_0} \tau^i\ \Phi_i(t,\ x) + \tau^{r-p/2}\ \Phi_{r-p/2,\,\tau}(t,\ x),$$

$$(t,\ x) \in Q_h^\tau,\quad s_0 = [r - p/2 - 1/2].$$

Les fonctions Φ_i, $i = 1, \ldots, s_0$, sont celles du théorème 4.6, et la fonction discrète $\Phi_{r-p/2,\tau}$ est bornée pour la métrique uniforme:

$$\max_{\overline{Q}_h^\tau} |\Phi_{r-p/2,\tau}(t,\ x)| \leqslant$$

$$\leqslant \max_{\overline{Q}_h^\tau} |\Phi_{r,\tau}(t,\ x)|\, h^{p/2} + \sum_{i=s_0+1}^{r} h^{i-s_0} \max_{\overline{Q}_h^\tau} |\Phi_i(t,\ x)|.$$

Les fonctions Φ_i étant continues sur $\overline{Q}$ sont uniformément bornées par une constante commune c_{11}, et l'estimation uniforme de $\Phi_{r,\tau}$ découle de l'inégalité

$$\max_{x\in\overline{\Omega}_h} |\Phi_{r,\tau}(t,\ x)| \leqslant \frac{1}{h^{p/2}} \|\Phi_{r,\tau}\|$$

vu que

$$\|\Phi_{r,\tau}\| \leqslant c_6 \quad \forall\, t \in \bar{\omega}_\tau$$

On réunit les inégalités obtenues, il vient

$$\max_{\overline{Q}_h^\tau} |\Phi_{r-p/2,\tau}(t,\ x) \leqslant c_6 + \sum_{i=s_0+1}^{r} c_{11} \leqslant c_6 + (r-1)\, c_{11},$$

i.e. on a un analogue du théorème 4.7 pour la métrique uniforme

THÉORÈME 4.8. *On suppose vérifiées les conditions* (4.3) *pour le problème* (4.1). (4.2) *avec un certain* $r \geqslant p/2 + 1$. *Si* M_i *vérifient les inégalités*

$$\frac{M_{i+1}}{M_i} \geqslant 1 + c_{12}, \qquad i = 1, \ldots, s_1; \qquad s_1 = [r - p/2 - 1/2] - 1,$$

et si γ_i *satisfont au système*

$$\sum_{i=1}^{s_1+1} \gamma_i = 1,$$

$$\sum_{i=1}^{s_1+1} \gamma_i \frac{1}{M_i^l} = 0, \; l = 1, \ldots, s_1,$$

alors γ_i *sont bornés, et la solution* $\overline{\Phi}$ *de la forme* (4.13) *est évaluée par*

$$\max_{\omega_\tau \times \overline{\Omega}_h} |\Phi(t, \mathbf{x}) - \overline{\Phi}(t, \mathbf{x})| \leqslant \sum_{i=1}^{s_1+1} |\gamma_i| \tau_i^{r-p/2} c_{11}.$$

où c_{11} *est indépendante de* τ *et* h.

CHAPITRE 6

EXTRAPOLATION DANS LA MÉTHODE DE RÉGULARISATION

Les problèmes relevant de la résolution numérique des équations différentielles ne sont pas les seuls à être abordés par l'extrapolation sur un paramètre. Quitte à la modifier de telle ou telle façon, on l'applique à d'autres problèmes d'Analyse numérique.

L'extrapolation sur des paramètres petits s'avère particulièrement intéressante. S'agissant des problèmes différentiels, ce paramètre a été le pas de discrétisation du schéma aux différences approchant le problème initial. Des cas peuvent cependant se présenter où l'on voit intervenir en outre d'autres paramètres indépendants qu'on fait tendre à la limite afin de raffiner la solution. La méthode de régularisation pour les systèmes algébriques dégénérés est du nombre.

La régularisation s'applique également aux équations différentielles ordinaires ou intégrales dont les valeurs propres forment un spectre continu ou discret à point de condensation $\lambda = 0$. On les réduit au préalable à des problèmes d'algèbre linéaire, ce qui donne de règle deux paramètres petits : le pas du réseau et le paramètre de régularisation, et l'on atteint une grande précision à condition de les faire tendre vers 0. La théorie de la régularisation est développée par A. Tikhonov, M. Lavrentiev, V. Ivanov, J.-L. Lions, V. Morozov et d'autres savants.

Dans ce chapitre nous nous occuperons uniquement des cas où l'on extrapole les solutions des systèmes algébriques linéaires sur le paramètre de régularisation. S'agissant des problèmes plus compliqués (tels les problèmes différentiels avec régularisation), on utilise en outre les méthodes décrites dans les chapitres précédents.

6.1. Régularisation d'un système dégénéré d'équations algébriques linéaires

Soit $\mathbf{C}^n$ l'espace hermitien de dimension n. On définit sur $\mathbf{C}^n$ le produit scalaire et la norme*

$$(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} u_i \bar{v}_i, \quad \| \mathbf{u} \| = (\mathbf{u}, \mathbf{u})^{1/2}.$$

* Les indices inférieurs des vecteurs sont les numéros des composantes.

On considère le système d'équations algébriques linéaires

$$Ax = \mathbf{f} \tag{1.1}$$

de matrice complexe dégénérée A de dimension $n \times n$, les vecteurs $\mathbf{x}$ et $\mathbf{f}$ étant dans $\mathbf{C}^n$.

On introduit la notation

$$U = \{ \mathbf{u} \in \mathbf{C}^n : \| A\mathbf{u} - \mathbf{f} \| = \inf_{\mathbf{v} \in \mathbf{C}^n} \| A\mathbf{v} - \mathbf{f} \| \}.$$

Un élément $\mathbf{u}^f \in U$ s'appelle *pseudo-solution normale* du système (1.1) si

$$\| \mathbf{u}^f \| = \min_{\mathbf{u} \in U} \| \mathbf{u} \| . \tag{1.2}$$

Ces conditions définissent le vecteur $\mathbf{u}^f$ de façon unique (voir [136]). Nous allons décrire un algorithme de construction de la pseudo-solution normale. On prémultiplie (1.1) par la matrice hermitienne conjuguée de A :

$$A^* A\mathbf{x} = A^*\mathbf{f}. \tag{1.3}$$

Soit une matrice unitaire P et une matrice diagonale $\Lambda = \operatorname{diag} \{ \lambda_1, \lambda_2, \ldots, \lambda_n \}$ telles que

$$A^* A = P \Lambda P^*. \tag{1.4}$$

On passe dans (1.3) aux variables $\mathbf{y} = P^*\mathbf{x}$:

$$\Lambda \mathbf{y} = P^* A^* \mathbf{f}. \tag{1.5}$$

Si l'on pose $\mathbf{F} = P^* A^* \mathbf{f}$, la pseudo-solution normale est définie par la formule

$$\mathbf{u}^f = P\mathbf{y}, \tag{1.6}$$

où le vecteur $\mathbf{y}$ a pour composantes

$$y_i = \begin{cases} F_i/\lambda_i & \text{si} \quad \lambda_i \neq 0, \\ 0 & \text{si} \quad \lambda_i = 0. \end{cases} \tag{1.7}$$

On approche le problème initial de minimiser la fonctionnelle sur U par un autre problème qui consiste à chercher le minimum sur $\mathbf{C}^n$ tout entier (voir [134], [136]).

La méthode de régularisation s'énonce comme suit: trouver dans $\mathbf{C}^n$ le point de minimum de la fonctionnelle

$$\| \mathbf{u} \|^2 + \frac{1}{\varepsilon} \| A\mathbf{u} - \mathbf{f} \|^2, \tag{1.8}$$

ε étant un paramètre positif. Le problème admet une solution unique $\mathbf{u}^\varepsilon$, et

$$\| \mathbf{u}^j - \mathbf{u}^\varepsilon \| \to 0 \text{ pour } \varepsilon \to 0.$$

La dérivabilité de (1.8) entraîne la condition de minimum

$$B^\varepsilon \mathbf{u}^\varepsilon = \mathbf{H}, \tag{1.9}$$

avec $B^\varepsilon = A^*A + \varepsilon I$, $\mathbf{H} = A^*\mathbf{f}$ (I étant une matrice unité).

On suppose que P et Λ sont celles de (1.4), si bien que le problème (1.9) se ramène par la substitution $\mathbf{y}^\varepsilon = P^* \mathbf{u}^\varepsilon$ au système d'équations de matrice diagonale

$$(\Lambda + \varepsilon I) \mathbf{y}^\varepsilon = \mathbf{F}. \tag{1.10}$$

On rappelle que $\mathbf{F} = P^* A^* \mathbf{f}$ et que $\lambda_i \geqslant 0$ pour tout $i = 1, 2, \ldots, n$. Etant donnée la dernière condition, le système (1.10) est possible, et sa solution unique est définie par les égalités [136]

$$y_i^\varepsilon = \frac{F_i}{\lambda_i + \varepsilon}, \qquad i = 1, \ldots, n. \tag{1.11}$$

On se propose d'approcher $\mathbf{u}^\varepsilon$ par une somme de solutions associées à d'autres paramètres de régularisation. On signale le grand rôle qui incombe dans les relations (1.11) à la fonction

$$\gamma(\varepsilon) = \frac{1}{\lambda + \varepsilon}. \tag{1.12}$$

Quand $\lambda > 0$, elle admet des dérivées de tous ordres pour tout ε non négatif, et on arrive sans peine à les évaluer:

$$|\gamma^k|(\varepsilon)| \leqslant \frac{k!}{\lambda^{k+1}}. \tag{1.13}$$

Soit $\varepsilon_1 > \varepsilon_2 > \ldots > \varepsilon_{l+1} \geqslant 0$ une suite des valeurs décroissantes de ε. On utilise le polynôme de Lagrange:

$$\gamma(\varepsilon_{l+1}) = \sum_{i=1}^{l} \alpha_i \gamma(\varepsilon_i) + Q(\varepsilon_{l+1}), \tag{1.14}$$

où

$$\alpha_i = \prod_{\substack{j=1 \\ j \neq i}}^{l} \frac{\varepsilon_{l+1} - \varepsilon_j}{\varepsilon_i - \varepsilon_j}. \tag{1.15}$$

On tire de [67] l'estimation du reste:

$$| Q(\varepsilon_{l+1}) | \leqslant \frac{1}{\lambda^{l+1}} \prod_{j=1}^{l} \varepsilon_j .$$

Si $\lambda_i \neq 0$, on a donc par suite de (1.11)

$$y_i^{\varepsilon_{l+1}} = \sum_{j+1}^{l} \alpha_j y_i^{\varepsilon_j} + q_i (\varepsilon_{l+1}), \tag{1.16}$$

où

$$| q_i (\varepsilon_{l+1}) | \leqslant F_i \lambda_i^{-l-1} \prod_{j=1}^{l} \varepsilon_j .$$

Si $\lambda_i = 0$, la fonction $\gamma(\varepsilon)$ n'est pas régulière au point 0, et les raisonnements ci-dessus n'aboutissent pas. Mais l'orthogonalité du vecteur $\mathbf{H} = A^*\mathbf{f}$ au noyau de la matrice $A^* A$ implique dans ce cas l'égalité $F_i = 0$. On désigne par λ_0 la plus petite valeur λ_i non nulle.

La relation définitive, conséquence de (1.16) (pour $\lambda_i \neq 0$) et de $F_i = 0$ (pour $\lambda_i = 0$), s'écrit

$$\mathbf{y}^{\varepsilon_{l+1}} = \sum_{j=1}^{l} \alpha_j \mathbf{y}^{\varepsilon_j} + \mathbf{q} (\varepsilon_{l+1}), \tag{1.17}$$

où

$$\| \mathbf{q} \| \leqslant \lambda_0^{-l-1} \prod_{j=1}^{l} \varepsilon_j \| \mathbf{F} \| .$$

Etant donné le développement (1.17), on démontre le résultat suivant.

LEMME 1.1. *Les solutions des problèmes régularisés* (1.9) *satisfont à la relation*

$$\mathbf{u}^{\varepsilon_{l+1}} = \sum_{j=1}^{l} \alpha_j \mathbf{u}^{\varepsilon_j} + \mathbf{r} (\varepsilon_{l+1}) \tag{1.18}$$

où

$$\| \mathbf{r} \| \leqslant c_1 \prod_{j=1}^{l} \varepsilon_j . \tag{1.19}$$

c_1 étant indépendante de ε_j.

DÉMONSTRATION. On prémultiplie (1.17) par la matrice P, il vient la formule (1.18), avec $\mathbf{r} = P\mathbf{q}$. On obtient l'estimation (1.19) en recourant deux fois à la propriété « la multiplication par une matrice unitaire conserve la norme », i.e.

$$\| \mathbf{H} \| = \| P\mathbf{F} \| = \| \mathbf{F} \|, \quad \| \mathbf{r} \| = \| P\mathbf{q} \| = \| \mathbf{q} \|$$

La sensibilité de la solution des problèmes (1.9) aux variations du second membre constitue un problème du plus haut intérêt. La nécessité d'approcher du second membre exact de (1.9) tient à ce que, *primo*, on calcule dans la pratique $A^*\mathbf{f}$ avec une erreur même si le vecteur $\mathbf{f}$ est connu exactement, et, *secundo*, le système (1.9) est résolu approximativement, i.e. on cherche un vecteur $\check{\mathbf{u}}$ tel que le résidu $B^\varepsilon \check{\mathbf{u}} - \mathbf{H}$ soit suffisamment petit sans être nul. On cherche donc la solution exacte du système

$$B^\varepsilon \mathbf{v}^\varepsilon = \mathbf{H}^\delta, \tag{1.20}$$

le vecteur inconnu $\mathbf{H}^\delta$ vérifiant l'inégalité

$$\| \mathbf{H}^\delta - \mathbf{H} \| \leqslant \delta, \tag{1.21}$$

avec $\delta > 0$ petit. Il est évident que $B^\varepsilon (\mathbf{v}^\varepsilon - \mathbf{u}^\varepsilon) = \mathbf{H}^\delta - \mathbf{H}$. Comme la plus petite valeur propre de B^ε vaut ε, l'estimation (1.21) entraîne de suite

$$\| \mathbf{v}^\varepsilon - \mathbf{u}^\varepsilon \| \leqslant \frac{\delta}{\varepsilon}. \tag{1.22}$$

Les estimations obtenues nous autorisent à proposer un procédé pour raffiner la solution des problèmes régularisés.

Soit $\varepsilon_1 > \varepsilon_2 > \varepsilon_3 > \ldots > \varepsilon_k > 0$ une suite de paramètres de régularisation pour lesquels on a trouvé la solution des problèmes

$$B^{\varepsilon_i} \mathbf{v}^{\varepsilon_i} = \mathbf{H}^{\delta_i}. \tag{1.23}$$

avec les matrices $B^{\varepsilon_i} = (A^* A + \varepsilon_i I)$ définies positives et $\mathbf{H}^{\delta_i} \in \mathbf{C}^n$, et

$$\| \mathbf{H}^{\delta_i} - \mathbf{H} \| \leqslant \delta_i. \tag{1.24}$$

La solution améliorée est construite moyennant la formule

$$\mathbf{w}^k = \sum_{j=1}^{k} \alpha_j \mathbf{v}^{\varepsilon_j}, \tag{1.25}$$

où les poids

$$\alpha_j = \prod_{\substack{i=1 \\ i \neq j}}^{k} \frac{-\varepsilon_i}{\varepsilon_j - \varepsilon_i}.$$

THÉORÈME 1.2. *Soit* $\mathbf{u}^j$ *la pseudo-solution normale du système* (1.1) *et* $\mathbf{w}^k$ *la solution améliorée définie par* (1.25). *La différence* $\mathbf{w}^k - \mathbf{u}^j$ *est majorée par*

$$\| \mathbf{w}^k - \mathbf{u}^j \| \leqslant \sum_{j=1}^{k} | \alpha_j | \frac{\delta_j}{\varepsilon_j} + c_2 \prod_{j=1}^{k} \varepsilon_j. \tag{1.26}$$

où la constante c_2 *est indépendante de* ε_j.

DÉMONSTRATION. Puisque $\mathbf{u}^f = \mathbf{u}^\varepsilon$ pour $\varepsilon = 0$, le lemme 1.1 entraîne pour $l = k$ et $\varepsilon_{k+1} = 0$

$$\mathbf{u}^f - \sum_{j=1}^{k} \alpha_j \mathbf{u}^{\varepsilon_j} = \mathbf{r}(0),$$

avec

$$\| \mathbf{r}(0) \| \leqslant c_2 \prod_{j=1} \varepsilon_j.$$

Si l'on utilise l'estimation

$$\| \mathbf{u}^{\varepsilon_j} - \mathbf{v}^{\varepsilon_j} \| \leqslant \delta_j / \varepsilon_j,$$

on a

$$\| \mathbf{u}^f - \mathbf{w}^k \| \leqslant \| \mathbf{u}^f - \sum_{j=1}^{k} \alpha_j \mathbf{u}^{\varepsilon_j} \| + \| \sum_{j=1}^{k} \alpha_j (\mathbf{u}^{\varepsilon_j} - \mathbf{v}^{\varepsilon_j} \| \leqslant$$

$$\leqslant \| \mathbf{r}(0) \| + \sum_{j=1}^{k} | \alpha_j | \, \| \mathbf{u}^{\varepsilon_j} - \mathbf{v}^{\varepsilon_j} \|,$$

d'où le résultat désiré.

Quant aux quantités figurant dans (1.26), certains éclaircissements s'imposent. Les coefficients α_j dépendent en général du choix de ε_j. Mais on peut prendre ε_j tels que $| \alpha_j |$ restent bornés pour tous les ε_j tendant vers 0. On partage par exemple le segment $[0, \varepsilon_1]$ par les points $\varepsilon_2, \ldots, \varepsilon_k$ régulièrement espacés. Dans ce cas, $| \alpha_j |$ sont en général indépendants de ε_j. Un procédé plus général consiste à choisir

$$\frac{\varepsilon_j}{\varepsilon_{j+1}} \geqslant c_3 > 1, \; j = 1, \ldots, k-1, \tag{1.27}$$

la constante c_3 ne dépendant pas de ε_j, auquel cas le lemme 2.3, § 7.2 entraîne

$$| \alpha_j | \leqslant \left(\frac{c_3}{c_3 - 1} \right)^k$$

quel que soit $j = 1, \ldots, k$. La sensibilité de δ_j à la diminution de ε_j est fonction du procédé de recherche de la solution des problèmes (1.23). Une étude plus poussée montre cependant que si la méthode itérative consiste à obtenir le résidu le plus petit, alors δ_j varient peu. Par contre, cela exige de nombreuses itérations. Nous décrirons plus loin un artifice qui vient à bout de ce défaut.

On diminue la quantité de calcul à exécuter pour des paramètres de régularisation petits à l'aide de l'algorithme ci-dessous qui permet d'utiliser au mieux la régularité des solutions par rapport à ε.

Supposons choisis les paramètres $\varepsilon_1 > \varepsilon_2 > \ldots > \varepsilon_k > 0$, et soit le problème

$$(A^* A + \varepsilon_1 I)\, \mathbf{u}^{\varepsilon_1} = A^* \mathbf{f}. \tag{1.28}$$

On admet que le problème a pour solution approchée $\mathbf{v}^{\varepsilon_1}$ et que le résidu relatif à $\mathbf{v}^{\varepsilon_1}$ est égal en norme à δ_1.

Le problème suivant est

$$(A^* A + \varepsilon_2 I)\, \mathbf{u}^{\varepsilon_2} = A^* \mathbf{f}.$$

On initialise avec le vecteur $\mathbf{v}^{\varepsilon_1}$ et on continue jusqu'à ce que les itérations aboutissent à une solution approchée $\mathbf{v}^{\varepsilon_2}$ telle que le résidu correspondant soit en norme de l'ordre de $\varepsilon_2 \delta_1/\varepsilon_1$ pour que les contributions de l'erreur sur $\mathbf{v}^{\varepsilon_1}$ et de celle sur $\mathbf{v}^{\varepsilon_2}$ à la solution définitive soient de même ordre.

Conformément au lemme, l'approximation initiale $\mathbf{w}^3$ du troisième problème

$$(A^* A + \varepsilon_3 I)\, \mathbf{u}^{\varepsilon_3} = A^* \mathbf{f}$$

est définie par la formule

$$\mathbf{w}^3 = \frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_1 - \varepsilon_2} \mathbf{v}^{\varepsilon_1} + \frac{\varepsilon_3 - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} \mathbf{v}^{\varepsilon_2}.$$

Le procédé itératif est arrêté après le pas donnant le résidu en $\varepsilon_3 \delta_1/\varepsilon_1$ pour la norme.

Ainsi, on résout approximativement le i-ième problème

$$(A^* A + \varepsilon_i I)\, \mathbf{u}^{\varepsilon_i} = A^* \mathbf{f}$$

tant qu'on n'obtient pas le résidu en $\varepsilon_i \delta_1/\varepsilon_1$ pour la norme, et on initialise avec le vecteur

$$\mathbf{w}_i = \sum_{j=1}^{i-1} \beta_j \mathbf{v}^{\varepsilon_j},$$

où

$$\beta_j = \prod_{\substack{l=1 \\ l \neq j}}^{i-1} \frac{\varepsilon_i - \varepsilon_l}{\varepsilon_j - \varepsilon_l}.$$

et $\mathbf{v}^{\varepsilon_j}$ sont déjà connus.

Dès qu'on résout le k-ième problème, on utilise la formule (1.25) pour construire la solution améliorée. Cette solution vérifie le théorème 1.2.

Remarque. Le cas réel est traité de façon analogue.

Exemple. Soit le problème (1.1), où la matrice A et le vecteur $\mathbf{f}$ sont réels et prennent les valeurs du tableau 6.1. On connaît les

Tableau 6.1

Matrice A et second membre f

−0,954	−0,028	−0,014	0,102	0,078	0,130	0,030	−0,330	−0,054	0,0	1,0
−0,028	0,704	−0,148	0,564	−0,204	−0,340	−0,540	−0,06	−0,228	0,0	2,0
−0,014	−0,148	0,926	0,282	−0,102	−0,170	−0,270	−0,03	−0,114	0,0	3,0
0,102	−0,564	0,282	1,974	0,486	0,810	1,110	−0,21	0,402	0,0	4,0
0,078	−0,204	−0,102	0,486	0,054	−0,090	−0,210	−0,690	−0,222	0,0	5,0
0,130	−0,340	−0,170	0,810	0,090	0,150	−0,350	−1,150	−0,370	0,0	6,0
0,030	−0,540	−0,270	1,110	0,210	−0,350	0,150	−0,650	−0,470	0,0	7,0
−0,330	−0,06	−0,03	−0,21	−0,690	−1,150	−0,650	0,150	0,170	0,0	8,0
−0,054	−0,228	−0,114	0,402	−0,222	−0,370	−0,470	0,170	3,84	0,0	9,0
0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	10,0

valeurs propres de A, à savoir —2, —1, 0, 0, 0, 1, 2, 2, 3, 4. La dernière ligne du tableau révèle l'incompatibilité du système. On résout (1.9) pour les paramètres $\varepsilon_1 = 0{,}01$, $\varepsilon_2 = 0{,}5 \cdot 10^{-2}$, $\varepsilon_3 = 1/3 \cdot 10^{-2}$, et on construit moyennant (1.25) le correcteur linéaire $\mathbf{w}^3$ à partir des solutions $\mathbf{u}^{\varepsilon_1}$, $\mathbf{u}^{\varepsilon_2}$, $\mathbf{u}^{\varepsilon_3}$ ainsi trouvées. Le tableau 6.2 donne la pseudo-solution normale $\mathbf{u}^f$ et les erreurs $\mathbf{u}^f - \mathbf{u}^{\varepsilon_i}$, $\mathbf{u}^f - \mathbf{w}^3$, les erreurs d'arrondi altérant au plus le dernier chiffre significatif des résultats.

Tableau 6.2

Numéro de la composante	$\mathbf{u}^f$	$\mathbf{u}^f - \mathbf{u}^{\varepsilon_1}$	$\mathbf{u}^f - \mathbf{u}^{\varepsilon_2}$	$\mathbf{u}^f - \mathbf{u}^{\varepsilon_3}$	$\mathbf{u}^f - \mathbf{w}^3$
1	−0,9045	$-4{,}9 \cdot 10^{-3}$	$-3{,}3 \cdot 10^{-3}$	$-1{,}6 \cdot 10^{-3}$	$-7{,}8 \cdot 10^{-8}$
2	−1,5090	$-7{,}6 \cdot 10^{-3}$	$-5{,}1 \cdot 10^{-3}$	$-2{,}5 \cdot 10^{-3}$	$-1{,}1 \cdot 10^{-7}$
3	0,2455	$-1{,}3 \cdot 10^{-3}$	$-8{,}9 \cdot 10^{-4}$	$-4{,}5 \cdot 10^{-4}$	$-5{,}2 \cdot 10^{-8}$
4	4,6468	$1{,}4 \cdot 10^{-2}$	$9{,}5 \cdot 10^{-3}$	$4{,}7 \cdot 10^{-3}$	$1{,}7 \cdot 10^{-7}$
5	−2,4135	$-1{,}1 \cdot 10^{-2}$	$-7{,}9 \cdot 10^{-3}$	$-3{,}9 \cdot 10^{-3}$	$-1{,}7 \cdot 10^{-7}$
6	−4,0225	$-1{,}9 \cdot 10^{-2}$	$-1{,}3 \cdot 10^{-2}$	$-6{,}6 \cdot 10^{-3}$	$-2{,}8 \cdot 10^{-7}$
7	−1,5225	$4{,}9 \cdot 10^{-3}$	$3{,}3 \cdot 10^{-2}$	$1{,}7 \cdot 10^{-2}$	$2{,}6 \cdot 10^{-7}$
8	−2,2275	$4{,}1 \cdot 10^{-3}$	$2{,}8 \cdot 10^{-3}$	$1{,}4 \cdot 10^{-3}$	$2{,}6 \cdot 10^{-7}$
9	1,2205	$-2{,}6 \cdot 10^{-3}$	$-1{,}8 \cdot 10^{-3}$	$-9{,}0 \cdot 10^{-4}$	$-5{,}6 \cdot 10^{-8}$
10	0,0	0,0	0,0	0,0	0,0

Les résultats sont en accord parfait avec le théorème 1.2.

6.2. Régularisation des systèmes de matrice hermitienne

Soit le système d'équations algébriques linéaires

$$A\mathbf{x} = \mathbf{f}, \tag{2.1}$$

où la matrice A est hermitienne dégénérée d'ordre n et les vecteurs $\mathbf{x}$ et $\mathbf{f}$ sont dans $\mathbf{E}^n$.

On met A sous forme de produit

$$A = P\Lambda P^*, \tag{2.2}$$

P étant une matrice unitaire et $\Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ une matrice diagonale à éléments réels λ_i.

On passe dans (2.1) aux variables $\mathbf{y} = P^*\mathbf{x}$:

$$\Lambda\mathbf{y} = P^*\mathbf{f}.$$

et on pose $\mathbf{F} = P^*\mathbf{f}$, auquel cas la pseudo-solution normale de l'équation (2.1) est définie par la relation

$$\mathbf{x}^I = P\mathbf{y}. \tag{2.3}$$

avec $\mathbf{y} \in \mathbf{E}^n$ de composantes calculées par les formules

$$y_j = \begin{cases} F_j/\lambda_j & \text{si} \quad \lambda_j \neq 0, \\ 0 & \text{si} \quad \lambda_j = 0. \end{cases} \tag{2.4}$$

On aborde, au lieu de (2.1), le système suivant

$$(A + \mathbf{i}\varepsilon I)\,\mathbf{x}^\varepsilon = \mathbf{f}, \tag{2.5}$$

où ε est un paramètre réel petit, $\mathbf{i} = \sqrt{-1}$ et I est une matrice unité. Toutes les valeurs propres de la matrice $A + \mathbf{i}\varepsilon I$ sont de module supérieur à $|\varepsilon|$ et se situent dans le demi-plan $\operatorname{Im}\lambda > 0$ (cas $\varepsilon > 0$) ou $\operatorname{Im}\lambda < 0$ (cas $\varepsilon < 0$).

Remarque. S'agissant de A symétrique réelle et de $\mathbf{f}$ réel, le système complexe (2.5) est équivalent au système d'équations algébriques linéaires sur le corps des nombres réels

$$\begin{bmatrix} \varepsilon I & -A \\ A & \varepsilon I \end{bmatrix} \begin{bmatrix} \mathbf{v}^\varepsilon \\ \mathbf{w}^\varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{f} \end{bmatrix}. \tag{2.6}$$

Ceci étant, entre les solutions a lieu la relation $\mathbf{x}^\varepsilon = \mathbf{v}^\varepsilon - \mathbf{i}\mathbf{w}^\varepsilon$. On note que selon que $\varepsilon > 0$ ou $\varepsilon < 0$, la matrice de (2.6) est définie positive ou définie négative dans l'espace euclidien (réel) de dimension $2n$ muni du produit scalaire

$$(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{2n} a_i b_i.$$

Le changement de variables $\mathbf{z}^\varepsilon = P^*\mathbf{x}^\varepsilon$ ramène le système (2.5) à

$$(\Lambda + \mathbf{i}\varepsilon I)\,\mathbf{z}^\varepsilon = P^*\mathbf{f}.$$

On approche la pseudo-solution normale par la somme de plusieurs solutions régularisées obtenues pour divers paramètres de régularisation.

Soit $k > 1$ un entier. Étant données les formules (2.4), on constate aisément que le problème d'extrapolation dépend du comportement de la fonction

$$\gamma_\lambda(\varepsilon) = \frac{1}{\lambda + \mathbf{i}\varepsilon}$$

pour λ et ε réels. Si $\lambda \neq 0$, la fonction $\gamma_\lambda(\varepsilon)$ est indéfiniment dérivable par rapport à ε, et on évalue facilement ses dérivées en valeur absolue :

$$|\gamma_\lambda^k| \leqslant \frac{k!}{|\lambda|^{k+1}}.$$

On développe $\gamma_\lambda(\varepsilon)$ en formule de Taylor et on s'arête après k premiers termes, il vient

$$\gamma_\lambda(\varepsilon) = \sum_{j=0}^{k-1} \frac{\varepsilon^j}{\lambda^{j+1}} (-i)^j + \zeta(\varepsilon), \qquad (2.7)$$

où $|\zeta(\varepsilon)| \leqslant |\varepsilon|^k / |\lambda|^{k+1}$. Soit maintenant ε_j, $j=1, \ldots, k$, un jeu de paramètres non nuls. On forme la combinaison linéaire

$$\sum_{j=1}^{k} \alpha_j \gamma_\lambda(\varepsilon_j), \qquad (2.8)$$

avec les poids α_j tels qu'on approche au mieux $\gamma_\lambda(0)$ pour $\varepsilon_j \to 0$. On additionne les développements (2.7) et on annule k premiers coefficients des puissances de λ, il vient

$$\sum_{j=1}^{k} \alpha_j = 1,$$
$$\sum_{j=1}^{k} \alpha_j \varepsilon_j^l = 0, \quad l=1, \ldots, k-1. \qquad (2.9)$$

La solution de ce système est

$$\alpha_j = \prod_{\substack{l=1 \\ l \neq j}}^{k} \frac{-\varepsilon_l}{\varepsilon_j - \varepsilon_l}. \qquad (2.10)$$

Avec cette famille de poids α_j, la combinaison linéaire (2.8) approche $\gamma_\lambda(0)$ avec la précision (voir [67])

$$\left| \gamma_\lambda(0) - \sum_{j=1}^{k} \alpha_j \gamma_\lambda(\varepsilon_j) \right| \leqslant \frac{1}{|\lambda|^{k+1}} \prod_{j=1}^{k} \varepsilon_j.$$

Si $\lambda_l \neq 0$, on a donc la relation

$$y_l = z_0^l = \sum_{j=1}^{k} \alpha_j z_l^{\varepsilon_j} + q_l \prod_{j=1}^{k} \varepsilon_j. \qquad (2.11)$$

où

$$|q_i| \leqslant F_i \lambda_i^{-k-1}.$$

Dans le cas contraire, le système (2.1) est incompatible, si bien qu'il faut que

$$\sum_{j=1}^{k} \frac{\alpha_j}{\varepsilon_j} = 0. \tag{2.12}$$

D'où la relation

$$y_i = 0 = \sum_{j=1}^{k} \alpha_j \varepsilon_i^{\varepsilon_j}.$$

On garantit l'égalité (2.12) par deux procédés différents.

Le premier procédé consiste à remplacer (2.9) par le système

$$\begin{aligned} &\sum_{j=1}^{k} \alpha_j \varepsilon_j^{-1} = 0, \\ &\sum_{j=1}^{k} \alpha_j = 1, \\ &\sum_{j=1}^{k} \alpha_j \varepsilon_j^{l} = 0, \quad l = 1,\ 2,\ \ldots,\quad k-2, \end{aligned} \tag{2.13}$$

dont le déterminant est calculé sous forme explicite:

$$\prod_{1 \leqslant j < i \leqslant k} (\varepsilon_i - \varepsilon_j) \Big/ \prod_{i=1}^{k} \varepsilon_i.$$

Ce déterminant s'annule si $\varepsilon_i = \varepsilon_j$, $i \neq j$. Le procédé présente le désavantage de détériorer la précision. En effet, on a beau disposer de k paramètres, la dernière équation (2.9) n'est en général pas satisfaite, si bien qu'on extrapole en fait sur $k-1$ paramètres.

Dans le second procédé, on remplit la restriction supplémentaire (2.12) grâce à un choix spécial de ε_j.

LEMME 2.1. *Soit ε_j distincts deux à deux, non nuls et tels que*

$$\sum_{j=1}^{k} \varepsilon_j^{-1} = 0. \tag{2.14}$$

On a, pour les solutions α_j du système (2.9), *la relation*

$$\sum_{j=1}^{k} \alpha_j \varepsilon_j^{-1} = 0.$$

DÉMONSTRATION. Soit le système

$$\begin{bmatrix} 1 & \varepsilon_1 & \varepsilon_1^2 & \dots & \varepsilon_1^{k-1} \\ 1 & \varepsilon_2 & \varepsilon_2^2 & \dots & \varepsilon_2^{k-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \varepsilon_k & \varepsilon_k^2 & \dots & \varepsilon_k^{k-1} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_{k-1} \end{bmatrix} = \begin{bmatrix} \varepsilon_1^{-1} \\ \varepsilon_2^{-1} \\ \dots \\ \varepsilon_k^{-1} \end{bmatrix}. \tag{2.15}$$

Le déterminant de sa matrice est égal au déterminant de Vandermonde $V(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)$. Comme ε_i sont distincts deux à deux, le système possède donc une solution unique. On cherche b_0 par la règle de Cramer:

$$b_0 = \frac{1}{V(\varepsilon_1, \dots, \varepsilon_k)} \det \begin{vmatrix} \varepsilon_1^{-1} & \varepsilon_1 & \varepsilon_1^2 & \dots & \varepsilon_1^{k-1} \\ \varepsilon_2^{-1} & \varepsilon_2 & \varepsilon_2^2 & \dots & \varepsilon_2^{k-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \varepsilon_k^{-1} & \varepsilon_k & \varepsilon_k^2 & \dots & \varepsilon_k^{k-1} \end{vmatrix}.$$

Le lemme 2.7, § 7.2 explicite le dernier déterminant, si bien que

$$b_0 = \frac{1}{V(\varepsilon_1, \dots, \varepsilon_k)} \sum_{j=1}^{k} \varepsilon_j^{-1} V(\varepsilon_1, \dots, \varepsilon_{j-1}, \varepsilon_{j+1}, \dots, \varepsilon_k) = 0$$

en vertu de la condition (2.14).

On calcule de même b_j restants, $j = 1, \dots, k-1$, et

$$\sum_{i=1}^{k-1} \varepsilon_j^i b_i = \varepsilon_j^{-1}, \qquad j = 1, \dots, n. \tag{2.16}$$

On fait la somme des équations de (2.9) avec les poids b_l:

$$\sum_{l=1}^{k-1} b_l \sum_{j=1}^{k} \alpha_j \varepsilon_j^l = 0.$$

On change l'ordre de sommation et on utilise (2.16), il vient

$$\sum_{j=1}^{k} \alpha_j \sum_{l=1}^{k-1} b_l \varepsilon_j^l = \sum_{j=1}^{k} \alpha_j \varepsilon_j^{-1} = 0,$$

c.q.f.d.

Les résultats obtenus sont à la base d'une méthode de recherche de la pseudo-solution normale. Soit $k \geq 2$ un entirer et ε_j, $j = 1, \dots$ $\dots, k$, une suite de paramètres réels vérifiant (2.14). On suppose connues k solutions x^{ε_j} des systèmes régularisés (2.5) associés aux paramètres ε_j.

On forme la combinaison linéaire

$$\mathrm{x} = \sum_{j=1}^{k} \alpha_j \mathrm{x}^{\varepsilon_j}. \tag{2.17}$$

x_j étant solution du système (2.9). On a le

THÉORÈME 2.2. *Le vecteur* $\bar{\mathrm{x}}$ *défini par la formule* (2.17) *approche la pseudo-solution normale du problème* (2.1) *avec la précision relative*

$$\frac{\|\mathrm{x}^J - \bar{\mathrm{x}}\|}{\|\bar{\mathrm{x}}\|} \leqslant \mu^{-k} \prod_{j=1}^{k} |\varepsilon_j|. \tag{2.18}$$

μ étant la valeur propre non nulle de plus petit module de A et x^J *la pseudo-solution normale de* (2.1).

DÉMONSTRATION. On décompose les vecteurs x^J et $\bar{\mathrm{x}}$ dans la base formée de vecteurs propres de A. Avec les notations (2.2) à (2.4), on écrit le carré du numérateur du premier membre de (2.18):

$$\|\mathrm{x}^J - \mathrm{x}\|^2 = \sum_{l=1}^{n} f_l^2 \left| \frac{1}{\lambda_l} - \sum_{j=1}^{k} \alpha_j \frac{1}{\lambda_l + i\varepsilon_j} \right|^2,$$

les termes relatifs à $\lambda_l = 0$ étant omis. On se rappelle la formule (2.11) et on évalue chaque terme du second membre de la dernière relation:

$$f_l^2 \left| \frac{1}{\lambda_l} - \sum_{j=1}^{k} \frac{\alpha_j}{\lambda_l + i\varepsilon_j} \right|^2 \leqslant f_l^2 \lambda_l^{-2k-2} \left(\prod_{j=1}^{k} \varepsilon_j \right)^2, \quad \lambda_l \neq 0.$$

Ainsi,

$$\|\mathrm{x}^J - \mathrm{x}\|^2 \leqslant \mu^{-2k} \left(\prod_{j=1}^{k} \varepsilon_j \right)^2 \sum_{l=1}^{k} f_l^2 \lambda_l^{-2} = \mu^{-2k} \left(\prod_{j=1}^{k} \varepsilon_j \right)^2 \|\mathrm{x}\|^2,$$

qui entraîne de suite l'affirmation du théorème.

Comme on demande à la fois plusieurs solutions de (2.5) associées à ε différents, une remarque s'impose: on initialise chaque fois avec l'approximation obtenue à partir des approximations précédentes.

On suppose, par exemple, résolu le problème (2.5) avec ε_1 (on diminue le temps de calcul en prenant pour ε_1 celui des k paramètres connus qui est de plus grand module). On construit l'approximation initiale $\mathrm{x}^{\varepsilon_2, 0}$ pour le problème (2.5) associé à ε_2 à l'aide de la formule

$$\mathrm{x}^{\varepsilon_2, 0} = \frac{\varepsilon_1}{\varepsilon_2} \mathrm{x}^{\varepsilon_1}. \tag{2.19}$$

Ce choix garantit un coefficient convenable de la composante relative au noyau de la matrice A. En effet, une même composante intervient dans $\mathbf{x}^{\varepsilon_2}$ avec le poids $1/(i\varepsilon_2)$ et dans $\mathbf{x}^{\varepsilon_1}$ avec le poids $1/(i\varepsilon_1)$.

Connaissant les solutions $\mathbf{x}^{\varepsilon_1}$ et $\mathbf{x}^{\varepsilon_2}$, on forme l'approximation initiale

$$\mathbf{x}^{\varepsilon_3, 0} = \beta_1 \mathbf{x}^{\varepsilon_1} + \beta_2 \mathbf{x}^{\varepsilon_2},$$

les coefficients de pondération β_i étant obtenus moyennant le système

$$\frac{\beta_1}{\varepsilon_1} + \frac{\beta_2}{\varepsilon_2} = \frac{1}{\varepsilon_3},$$

$$\beta_1 + \beta_2 = 1.$$

Etant donnée la formule (2.11), on voit que $\mathbf{x}^{\varepsilon_3, 0}$ approche $\mathbf{x}^{\varepsilon_3}$ avec la précision $O(\bar{\varepsilon})$, où

$$\bar{\varepsilon} = \max_{1 \leqslant i \leqslant k} |\varepsilon_i|.$$

S'agissant de la solution $\mathbf{x}^{\varepsilon_4}$, l'approximation initiale est donnée par la formule

$$\mathbf{x}^{\varepsilon_4, 0} = \gamma_1 \mathbf{x}^{\varepsilon_1} + \gamma_2 \mathbf{x}^{\varepsilon_2} + \gamma_3 \mathbf{x}^{\varepsilon_3},$$

avec les poids définis à partir du système

$$\frac{\gamma_1}{\varepsilon_1} + \frac{\gamma_2}{\varepsilon_2} + \frac{\gamma_3}{\varepsilon_3} = \frac{1}{\varepsilon_4},$$

$$\gamma_1 + \gamma_2 + \gamma_3 = 1,$$

$$\varepsilon_1 \gamma_1 + \varepsilon_2 \gamma_2 + \varepsilon_3 \gamma_3 = \varepsilon_4,$$

et dans ce cas

$$\| \mathbf{x}^{\varepsilon_4, 0} - \mathbf{x}^{\varepsilon_4} \| = O(\bar{\varepsilon}^2).$$

On illustre les résultats obtenus par un exemple numérique où figurent la matrice A et le vecteur $\mathbf{f}$ du tableau 6.1. On résout le problème (2.5) pour trois valeurs du paramètre ε : $\varepsilon_1 = 0{,}01$, $\varepsilon_2 = 0{,}5 \cdot 10^{-2}$ et $\varepsilon_3 = -1/3 \cdot 10^{-2}$, ce qui garantit la condition (2.14) du lemme 2.1. Ci-dessous deux tableaux donnant la pseudo-solution normale, les erreurs sur les solutions régularisées $\mathbf{x}^{\varepsilon_i}$ et l'erreur sur la solution extrapolée $\bar{\mathbf{x}}$ (le tableau 6.3 donne les parties réelles de ces vecteurs et le tableau 6.4 leurs parties imaginaires).

Tableau 6.3

Numéro de la composante	$\mathrm{Re}\,(x^f)$	$\mathrm{Re}\,(x^f - x^{\varepsilon_1})$ $\varepsilon_1 = 0{,}01$	$\mathrm{Re}\,(x^f - x^{\varepsilon_2})$ $\varepsilon_2 = \varepsilon_1/2$	$\mathrm{Re}\,(x^f - x^{\varepsilon_3})$ $\varepsilon_3 = -\varepsilon_1/3$	$\mathrm{Re}\,(x^f - \bar{x})$
1	$-0{,}9045$	$-4{,}9\cdot 10^{-5}$	$-1{,}2\cdot 10^{-5}$	$-5{,}5\cdot 10^{-6}$	$6{,}5\cdot 10^{-9}$
2	$-1{,}5090$	$-7{,}7\cdot 10^{-5}$	$-1{,}9\cdot 10^{-5}$	$-8{,}5\cdot 10^{-6}$	$1{,}1\cdot 10^{-9}$
3	$0{,}2455$	$-1{,}3\cdot 10^{-5}$	$-3{,}3\cdot 10^{-6}$	$-1{,}5\cdot 10^{-6}$	$4{,}2\cdot 10^{-9}$
4	$4{,}6468$	$1{,}4\cdot 10^{-4}$	$-3{,}6\cdot 10^{-5}$	$1{,}6\cdot 10^{-5}$	$1{,}7\cdot 10^{-9}$
5	$-2{,}4135$	$-1{,}2\cdot 10^{-4}$	$-2{,}9\cdot 10^{-5}$	$-1{,}3\cdot 10^{-5}$	$-7{,}0\cdot 10^{-10}$
6	$-4{,}0225$	$-2{,}0\cdot 10^{-4}$	$-4{,}9\cdot 10^{-5}$	$-2{,}2\cdot 10^{-5}$	$1{,}9\cdot 10^{-9}$
7	$-1{,}5225$	$5{,}1\cdot 10^{-5}$	$1{,}2\cdot 10^{-5}$	$5{,}6\cdot 10^{-6}$	$2{,}1\cdot 10^{-9}$
8	$-2{,}2276$	$4{,}3\cdot 10^{-5}$	$1{,}0\cdot 10^{-5}$	$4{,}7\cdot 10^{-6}$	$1{,}9\cdot 10^{-9}$
9	$1{,}2205$	$-2{,}7\cdot 10^{-5}$	$-6{,}8\cdot 10^{-6}$	$-3{,}0\cdot 10^{-6}$	$5{,}3\cdot 10^{-10}$
10	$0{,}0$	$0{,}0$	$0{,}0$	$0{,}0$	$0{,}0$

Tableau 6.4

Numéro de la composante	$\mathrm{Im}\,(x^f)$	$\mathrm{Im}\,(x^f - x^{\varepsilon_1})$	$\mathrm{Im}\,(x^f - x^{\varepsilon_2})$	$\mathrm{Im}\,(x^f - x^{\varepsilon_3})$	$\mathrm{Im}\,(x^f - \bar{x})$
1	$0{,}0$	$-2{,}87\cdot 10^{1}$	$-5{,}75\cdot 10^{1}$	$8{,}63\cdot 10^{1}$	10^{-8}
2	$0{,}0$	$-5{,}75\cdot 10^{1}$	$-1{,}15\cdot 10^{2}$	$1{,}72\cdot 10^{2}$	10^{-8}
3	$0{,}0$	$-2{,}88\cdot 10^{1}$	$-5{,}75\cdot 10^{1}$	$8{,}63\cdot 10^{1}$	10^{-9}
4	$0{,}0$	$8{,}64\cdot 10^{1}$	$1{,}72\cdot 10^{2}$	$-2{,}59\cdot 10^{2}$	10^{-7}
5	$0{,}0$	$1{,}43\cdot 10^{2}$	$2{,}87\cdot 10^{2}$	$-4{,}30\cdot 10^{2}$	10^{-8}
6	$0{,}0$	$6{,}00$	$1{,}2\cdot 10^{1}$	$-1{,}8\cdot 10^{1}$	10^{-7}
7	$0{,}0$	$-1{,}43\cdot 10^{2}$	$-2{,}87\cdot 10^{2}$	$4{,}3\cdot 10^{2}$	10^{-7}
8	$0{,}0$	$1{,}44\cdot 10^{2}$	$2{,}88\cdot 10^{2}$	$-4{,}3\cdot 10^{2}$	10^{-7}
9	$0{,}0$	$-2{,}87\cdot 10^{1}$	$-5{,}75\cdot 10^{1}$	$8{,}6\cdot 10^{1}$	10^{-8}
10	$0{,}0$	10^{3}	$2{,}0\cdot 10^{3}$	$-3{,}0\cdot 10^{3}$	10^{-12}

Si l'on fait la comparaison avec les résultats du § 6.1, on constate que l'extrapolation basée sur la propriété de symétrie de la matrice A est plus efficace.

6.3. Extrapolation des solutions avec fonctions du type couche limite

Le fait d'utiliser un paramètre petit conduit souvent en physique mathématique à des résultats utiles et importants. On introduit par exemple couramment des dérivées d'ordre supérieur avec poids petit pour modifier les conditions aux limites ou le type de l'équation différentielle. Les méthodes d'extrapolation sont en l'occurrence d'un grand secours pour construire des algorithmes de calcul économiques. Nous allons examiner deux exemples dans lesquels l'algorithme d'extrapolation s'appuie sur certaines particularités des solutions.

Exemple 1. On remplace le problème initial

$$\begin{aligned} -\Delta u &= f \quad \text{dans} \quad \Omega, \\ u &= g \quad \text{sur} \quad \Gamma \end{aligned} \tag{3.1}$$

par

$$\begin{aligned} -\Delta u_\varepsilon &= f \quad \text{dans} \quad \Omega, \\ u_\varepsilon + \varepsilon \frac{\partial u_\varepsilon}{\partial n} &= g \quad \text{sur} \quad \Gamma. \end{aligned} \tag{3.2}$$

Ici Ω est un domaine borné bidimensionnel de frontière Γ régulière, $\partial u/\partial n$ la dérivée par rapport à la normale extérieure à Γ et $\varepsilon > 0$ un paramètre petit.

Ce changement de problème a par exemple pour but de passer à d'autres conditions aux limites (voir [33]), ce qui simplifie un peu la pratique de la méthode des éléments finis (voir [61], [115], [132]).

Avant d'aborder l'étude du développement asymptotique pour u_ε, on démontre pour le problème (3.2) le

Théorème 3.1. *Soit u_1 solution du problème*

$$\begin{aligned} -\Delta u_1 &= 0 \quad \textit{dans} \quad \Omega, \\ u_1 + \varepsilon \frac{\partial u_1}{\partial n} &= g_1 \quad \textit{sur} \quad \Gamma \end{aligned}$$

et u_2 solution du problème

$$\begin{aligned} -\Delta u_2 &= 0 \quad \textit{dans} \quad \Omega, \\ u_2 + \varepsilon \frac{\partial u_2}{\partial n} &= g_2 \quad \textit{sur} \quad \Gamma, \end{aligned}$$

avec $\Gamma \in C^{2+\lambda}$; $g_1, g_2 \in C^{1+\lambda}(\Gamma)$; $\lambda \in (0, 1)$.

L'inégalité

$$|g_1| < g_2 \quad sur \quad \Gamma$$

entraîne

$$|u_1| < u_2 \quad sur \quad \bar{\Omega}. \qquad (3.3)$$

DÉMONSTRATION. On pose le problème pour la différence $v = u_2 - u_1$:

$$-\Delta v = 0 \quad \text{dans} \quad \Omega,$$

$$v + \varepsilon \frac{\partial v}{\partial n} = g_2 - g_2 \quad \text{sur} \quad \Gamma.$$

On suppose que v prend une valeur négative dans $\bar{\Omega}$, auquel cas elle atteint par continuité sa plus petite valeur négative en un point $\mathbf{x}_0$. On montre que $\mathbf{x}_0 \in \Omega$. En effet, si $\mathbf{x}_0 \in \Gamma$, alors

$$\varepsilon \frac{\partial v}{\partial n}(\mathbf{x}_0) = g_2(\mathbf{x}_0) - g_1(\mathbf{x}_0) - v(\mathbf{x}_0) > 0;$$

la dérivée $\partial v/\partial n$ garde par continuité le signe dans l'intersection d'un voisinage de $\mathbf{x}_0$ et du domaine Ω. Aussi la fonction v est strictement décroissante le long de la normale intérieure à Γ en $\mathbf{x}_0$, ce qui contredit la condition de minimum en ce point. Donc $\mathbf{x}_0 \in \Omega$. Le principe du maximum entraîne alors $v \equiv$ const sur $\bar{\Omega}$.

Soit un point quelconque $\mathbf{x}$ de Γ. Comme $v =$ const, on a $\frac{\partial v}{\partial n}(\mathbf{x}) = 0$, si bien que la condition aux limites devient

$$v(\mathbf{x}) = g_2(\mathbf{x}) - g_1(\mathbf{x}).$$

On a supposé $v(\mathbf{x}) < 0$ (comme c'est le cas partout sur $\bar{\Omega}$), et $g_2(\mathbf{x}) - g_1(\mathbf{x}) \geqslant 0$ par hypothèse du théorème. Ainsi, l'hypothèse de v négative au moins en un point $\mathbf{x} \in \bar{\Omega}$ nous a conduit à une contradiction.

Donc

$$v = u_2 - u_1 > 0 \quad \text{sur} \quad \bar{\Omega}.$$

Si l'on pose $v = u_2 + u_1$, on obtient

$$u_2 + u_1 > 0 \quad \text{sur} \quad \bar{\Omega}.$$

Les deux dernières inégalités donnent (3.3), i.e. l'affirmation du théorème de comparaison 3.1.

Quelles sont les conditions sous lesquelles u_ε admet un développement asymptotique?

THÉORÈME 3.2. *On suppose que le problème* (3.1) *remplit les conditions de régularité*

$$\Gamma \in C^{l+2}, \quad g \in C^{l+2}(\Gamma), \; f \in C^l(\bar{\Omega}), \qquad (3.4)$$

avec l non entier. Sa solution admet le développement

$$u_\varepsilon = u + \sum_{k=1}^{s} \varepsilon^k v_k + \varepsilon^{s+1} w_\varepsilon \quad sur \quad \bar{\Omega}, \tag{3.5}$$

où $s = [l]$, les fonctions v_k ne dépendent pas de ε et

$$\max_{\bar{\Omega}} |w_\varepsilon| \leqslant c_1, \tag{3.6}$$

la constante c_1 étant indépendante de ε.

DÉMONSTRATION. On pose $v_0 = u$ et on définit v_i moyennant la suite de problèmes

$$\begin{gathered} -\Delta v_i = 0 \quad \text{dans} \quad \Omega, \\ v_i = -\frac{\partial v_{i-1}}{\partial n} \quad \text{sur} \quad \Gamma, \\ i = 1, \ldots, s. \end{gathered} \tag{3.7}$$

Aux termes du théorème 1.2, § 4.1, la solution u du problème (3.1) est dans $C^{l+2}(\bar{\Omega})$ et les fonctions $v_i \in C^{l+2-i}(\bar{\Omega})$ sont indépendantes de ε. On introduit la fonction

$$w_\varepsilon = \varepsilon^{-s-1} \left(u_\varepsilon - \sum_{i=0}^{s} \varepsilon_i v_i \right) \quad \text{sur} \quad \bar{\Omega}.$$

Il découle de (3.2)

$$\begin{gathered} -\Delta w_\varepsilon = 0 \quad \text{dans} \quad \Omega, \\ w_\varepsilon + \varepsilon \frac{\partial w_\varepsilon}{\partial n} = -\frac{\partial v_s}{\partial n} \quad \text{sur} \quad \Gamma. \end{gathered} \tag{3.8}$$

Puisque $v_s \in C^{l-s+2}(\bar{\Omega})$, on a $\partial v_s / \partial n \in C^{l-s+1}(\Gamma)$, si bien que la solution du problème (3.8) existe, est unique et appartient à $C^{l-s+2}(\bar{\Omega})$ (voir [26]). Les estimations de la solution et de ses dérivées dépendent en général de ε. On montre cependant que ce n'est pas toujours le cas en ce qui concerne w_ε. On applique le théorème 3.1 aux fonctions

$$w_\varepsilon, \; w = \max_\Gamma \left| \frac{\partial v_s}{\partial n} \right|.$$

Evidemment,

$$\begin{gathered} -\Delta w = 0 \quad \text{dans} \quad \Omega, \\ w + \varepsilon \frac{\partial w}{\partial n} = \max_\Gamma \left| \frac{\partial v_s}{\partial n} \right| \quad \text{sur} \quad \Gamma. \end{gathered}$$

Par conséquent,

$$|w_\varepsilon| \leqslant w \quad \text{sur} \quad \bar{\Omega}.$$

Si l'on pose

$$c_1 = \max_{\Gamma} \left| \frac{\partial v_\varepsilon}{\partial n} \right|,$$

on aboutit à (3.6), ce qui démontre le théorème.

Comme on cherche u et non u_ε, on utilise en vertu de (3.5) une extrapolation linéaire usuelle sur le paramètre ε.

On suppose qu'on est dans les hypothèses du théorème 3.2. On cherche $u_{\varepsilon/k}$, solutions associées à ε/k, $k = 1, \ldots, s+1$, des problèmes (3.2), et on forme le correcteur linéaire

$$U^E = \sum_{k=1}^{s+1} \gamma_k \, u_{\varepsilon/k} \quad \text{sur} \quad \bar{\Omega}, \tag{3.9}$$

ou

$$\gamma_k = \frac{(-1)^{s-k+1} \, k^{s+1}}{k!(s-k+1)!}.$$

THÉORÈME 3.3. *On suppose vérifiées les hypothèses du théorème 3.2. Le correcteur* (3.9) *admet l'estimation*

$$|U^E - u| \leqslant c_2 \varepsilon^{s+1}, \tag{3.10}$$

avec la constante c_2 indépendante de ε.

DÉMONSTRATION. On utilise le développement (3.5) pour $k = 1, \ldots, s+1$, on forme le correcteur U^E et on applique le lemme 2.1, § 7.2, il vient

$$U^E = u + \varepsilon^{s+1} \sum_{k=1}^{s+1} \frac{\gamma_k}{k^{s+1}} w_{\varepsilon/k}.$$

On porte u à gauche, on prend le module des deux membres et on recourt à la majoration (3.6). On pose

$$c_2 = c_1 \sum_{k=1}^{s+1} \frac{|\gamma_k|}{k^{s+1}},$$

ce qui donne l'inégalité voulue (3.10)

Le résultat obtenu permet de calculer u avec une grande précision à partir de plusieurs solutions approchées $u_{\varepsilon/k}$ pour ε/k suffisamment importants. On conçoit que la caractéristique « économie »

des algorithmes de calcul s'en trouve sensiblement améliorée. On atteint une efficacité particulièrement grande si l'on accorde l'extrapolation sur ε avec l'extrapolation sur le pas du réseau de discrétisation.

EXEMPLE 2. On se place en dimension 1 pour simplifier l'exposé. Soit le problème modèle qui se pose dans le calcul du mouvement d'un milieu avec viscosité faible:

$$\begin{gathered} -\varepsilon^2 y_\varepsilon'' + a y_\varepsilon = f \quad \text{sur} \quad (0,\ 1), \\ y_\varepsilon(0) = y_0,\ y_\varepsilon(1) = y_1. \end{gathered} \tag{3.11}$$

Ici $\varepsilon > 0$ est un paramètre petit, $a(x) > 0$ et $f(x)$ sont des fonctions suffisamment régulières. D'après [5], on cherche la solution sous forme de somme

$$y_\varepsilon = u + b e^{-d/\varepsilon} + \varepsilon w_\varepsilon \quad \text{sur} \quad [0,\ 1], \tag{3.12}$$

u étant solution de:

$$au = f \quad \text{sur} \quad (0, 1), \tag{3.13}$$

problème qui est limite de (3.11). Les fonctions $b(x)$ et $d(x) \geqslant 0$ sont continues sur $[0,\ 1]$ et indépendantes de ε. Le reste w_ε admet la majoration

$$|w_\varepsilon| \leqslant c_3 \tag{3.14}$$

où la constante c_3 ne dépend pas de ε.

Dans le problème (3.11), on demande non la solution limite u (comme dans l'exemple 1), mais l'approximation y_{ε_0} pour un certain ε_0 suffisamment petit. On cherche cette approximation à l'aide de plusieurs solutions y_{ε_i}, ε_i étant sensiblement plus grands que ε_0.

Considérons le cas où l'on connaît la solution de (3.13). On trouve deux solutions y_{ε_1}, y_{ε_2} de (3.11) pour ε_1 et $\varepsilon_2 = \varepsilon_1/2$ supérieurs à ε_0. On néglige le reste dans le second membre de (3.12) et on passe aux formules approchées

$$y_{\varepsilon_1} \approx u + b e^{-d/\varepsilon_1},\ y_{\varepsilon_2} \approx u + b e^{-2d/\varepsilon_1}.$$

On en tire, pour les inconnues $b(x)$, $d(x)$, $x \in (0,\ 1)$, le système approché

$$b e^{-d/\varepsilon_1} \approx y_{\varepsilon_1} - u,\ b e^{-2d/\varepsilon_1} \approx y_{\varepsilon_2} - u.$$

On divise la première équation élevée au carré par la seconde, il vient

$$b \approx \frac{(y_{\varepsilon_1} - u)^2}{y_{\varepsilon_2} - u}.$$

Divisons la seconde équation par la première:

$$e^{-d/\varepsilon_1} \approx \frac{y_{\varepsilon_2} - u}{y_{\varepsilon_1} - u}.$$

On rappelle que

$$y_{\varepsilon_0} \approx u + be^{-d/\varepsilon_0}.$$

Aussi

$$y_{\varepsilon_0} \approx u + \frac{(y_{\varepsilon_1} - u)^2}{y_{\varepsilon_2} - u} \left(\frac{y_{\varepsilon_2} - u}{y_{\varepsilon_1} - u}\right)^{\varepsilon_1/\varepsilon_0}. \tag{3.15}$$

Comme le second membre renferme des quantités connues, on le prend tout naturellement pour une approximation de y_{ε_0}. On pose

$$Y_{\varepsilon_0} = u + (y_{\varepsilon_1} - u) \left(\frac{y_{\varepsilon_2} - u}{y_{\varepsilon_1} - u}\right)^{\varepsilon_1/\varepsilon_0 - 1}. \tag{3.16}$$

Cette formule n'est utile que pour $x \in [0, 1]$ tels que la fonction du type couche limite diffère de 0 au moins d'une quantité de l'ordre de ε_0:

$$|b(x)| \exp(-d(x)/\varepsilon_0) \geqslant c_4 \varepsilon_0, \tag{3.17}$$

avec c_4 une constante positive quelconque indépendante de ε. Si $b(0) \neq 0$, il existe au voisinage de 0 un segment $[0, \delta_1]$ sur lequel la condition (3.17) se trouve satisfaite. Si $b(1) \neq 0$, c'est une région autour de 1 qui renferme un tel segment, à savoir $[\delta_2, 1]$. Ces segments constituent un ensemble noté ω.

On se propose d'étudier la précision de la solution extrapolée Y_{ε_0} et la stabilité de (3.16) vis-à-vis des erreurs résultant du calcul approché de u, y_{ε_1}, y_{ε_2}. Soit $x \in \omega$ et les approximations respectives $\tilde{u}$, $\tilde{y}_{\varepsilon_1}$, $\tilde{y}_{\varepsilon_2}$ obtenues avec les erreurs

$$\alpha = \tilde{u}(x) - u(x),\ \alpha_{\varepsilon_1} = \tilde{y}_{\varepsilon_1}(x) - y_{\varepsilon_1}(x),$$
$$\alpha_{\varepsilon_2} = \tilde{y}_{\varepsilon_2}(x) - y_{\varepsilon_2}(x). \tag{3.18}$$

On utilise les valeurs approchées de (3.16), il vient la solution extrapolée

$$\tilde{Y}_{\varepsilon_0} = \tilde{u} + (\tilde{y}_{\varepsilon_1} - \tilde{u}) \left(\frac{\tilde{y}_{\varepsilon_2} - \tilde{u}}{\tilde{y}_{\varepsilon_1} - \tilde{u}}\right)^{\varepsilon_1/\varepsilon_0 - 1}. \tag{3.19}$$

Quelle en est la précision pour ε_1 petit et le rapport fixe

$$\varepsilon_1/\varepsilon_0 = k > 2? \tag{3 20}$$

THÉORÈME 3.4. *On suppose que la représentation* (3.12) *a lieu en un point* $x \in \omega$, *que le reste est évalué par* (3.14) *et que les erreurs* (3.18) *sont* $O(\varepsilon_1^l)$:

$$\max (|\alpha|, |\alpha_{\varepsilon_1}|, |\alpha_{\varepsilon_2}|) \leqslant c_5 \varepsilon_1^l, \text{ où } l > 1/k. \quad (3.21)$$

On a pour ε_1 *suffisamment petit*

$$|y_{\varepsilon_0}(x) - \tilde{Y}_{\varepsilon_0}(x)| \leqslant c_6 (\varepsilon_1 + |\alpha| + |\alpha_{\varepsilon_1}| + |\alpha_{\varepsilon_2}|). \quad (3.22)$$

DÉMONSTRATION. Il découle de (3.12) et (3.18)

$$\tilde{y}_{\varepsilon_1} - \tilde{u} = b \exp (d/\varepsilon_1) + \varepsilon_1 w_{\varepsilon_1} + \alpha_{\varepsilon_1} - \alpha,$$

$$\tilde{y}_{\varepsilon_2} - \tilde{u} = b \exp (-2d/\varepsilon_1) + \varepsilon_1 w_{\varepsilon_2}/2 + \alpha_{\varepsilon_2} - \alpha.$$

On désigne $\varepsilon_1 w_{\varepsilon_1} + \alpha_{\varepsilon_1} - \alpha$ par β_1; $\varepsilon_1 w_{\varepsilon_2}/2 + \alpha_{\varepsilon_2} - \alpha$ par β_2 et $\alpha - \varepsilon_0 w_{\varepsilon_0}$ par β_3, et on introduit la fonction

$$u(t_1, t_2, t_3) = u + \varepsilon_0 w_{\varepsilon_0} + t_3 + (b \exp (-d/\varepsilon_1) + t_1) \times$$
$$\times [(b \exp (-2d/\varepsilon_1) + t_2)/(b \exp (-d/\varepsilon_1) + t_1)]^{k-1}$$

On note que $u(0, 0, 0) = y_{\varepsilon_0}$ et $u(\beta_1, \beta_2, \beta_3) = \tilde{Y}_{\varepsilon_0}$. On démontre que u admet pour t_i suffisamment petits des dérivées bornées

$$\frac{\partial u}{\partial t_1}(t_1, t_2, t_3) = (2-k)[(b \exp(-2d/\varepsilon_1) + t_2)/(b \exp(-d/\varepsilon_1) + t_1)]^{k-1},$$

$$\frac{\partial u}{\partial t_2}(t_1, t_2, t_3) = (k-1)[b \exp(-2d/\varepsilon_1) + t_2)/(b \exp(-d/\varepsilon_1) + t_1)]^{k-2},$$

$$\frac{\partial u}{\partial t_3}(t_1, t_2, t_3) = 1.$$

On considère d'abord le dénominateur de

$$A = (b \exp (-2d/\varepsilon_1) + t_2) / (b \exp (-d/\varepsilon_1) + t_1).$$

L'estimation (3.17) entraîne

$$\exp (-d/\varepsilon_1) \geqslant (c_4 \varepsilon_0/|b|)^{\varepsilon_0/\varepsilon_1} = c_4 \varepsilon_1/|b_k|)^{1/k}.$$

Donc

$$|b| \exp (-d/\varepsilon_1) \geqslant c_7 \varepsilon_1^{1/k}. \quad (3.23)$$

Les majorations (3.14), (3.21) impliquent l'inégalité

$$|t_1| \leqslant |\beta_1| \leqslant c_3 \varepsilon_1 + 2\varepsilon_1^l c_5. \quad (3.24)$$

Si $l > 1$, alors (3.24) donne par suite de $\varepsilon_1 \leqslant 1$: $|t_1| \leqslant (c_3 + 2c_5) \varepsilon_1$. Si $l \leqslant 1$, alors le résultat correspondant est $|t_1| \leqslant (c_3 + 2c_5)\varepsilon_1^l$.

Dans les deux cas, $|t_1|$ est un infiniment petit d'ordre supérieur à celui du second membre de (3.23). Aussi

$$|b| \exp(-d/\varepsilon_1) \geqslant 2\,|t_1|$$
$$\forall \varepsilon_1 \leqslant \min\{1, [c_7/(c_3 + 2\,c_5)^2]^{k/(kl-1)}, [c_7/(c_3 + 2c_5)^2]^{k(k-1)}\}.$$

D'où

$$|b \exp(-d/\varepsilon_1) + t_1| \geqslant \frac{1}{2}|b| \exp(-d/\varepsilon_1).$$

On a, compte tenu de cette inégalité et de (3.23),

$$|A| \leqslant [|b| \exp(-2d/\varepsilon_1) + |t_2|]/\left[\frac{1}{2}|b| \exp(-d/\varepsilon_1)\right] \leqslant$$
$$\leqslant 2 \exp(-d/\varepsilon_1) + 2\varepsilon^{-1/k}\,|t_2|/c_7.$$

(3.14) et (3.21) entraînent

$$|t_2| \leqslant c_3\,\varepsilon_1/2 + 2c_5\,\varepsilon_1^l.$$

On réunit les deux inégalités et on se rappelle que $\varepsilon_1 < 1$, $d \geqslant 0$, $1/k < l$, il vient

$$|A| \leqslant 2 + (c_9 + 4c_5)/c_7 \equiv c_8.$$

Par conséquent,

$$\left|\frac{\partial u}{\partial t_i}(t_1, t_2, t_3)\right| \leqslant c_9 \forall i = 1, 2, 3, \forall t_i \in [0, \beta_i], \qquad (3.25)$$

où $c_9 \equiv \max\{1, (k-2)\,c_8^{k-1}, (k-1)\,c_8^{k-2}\}$.

On a par Lagrange

$$u(\beta_1, \beta_2, \beta_3) = u(0, 0, 0) + \beta_1 \frac{\partial u}{\partial t_1}(\xi_1, \beta_2, \beta_3) +$$
$$+ \beta_2 \frac{\partial u}{\partial t_2}(0, \xi_2, \beta_3) + \beta_3 \frac{\partial u}{\partial t_3}(0, 0, \xi_3), \text{ où } \xi_i \in [0, \beta_i].$$

On obtient donc moyennant (3.25)

$$|u(\beta_1, \beta_2, \beta_3) - u(0, 0, 0)| \leqslant c_9(|\beta_1| + |\beta_2| + |\beta_3|).$$

Le premier membre est égal à $|y_{\varepsilon_0}(x) - \tilde{Y}(x)|$. On évalue les termes à droite par (3.14), et on aboutit à (3.22), c.q.f.d.

L'exemple ci-dessous illustre l'efficacité de cet algorithme d'extrapolation. Soit, dans le problème (3.11), $a(x) = 1$, $f(x) = 10(2 - e^x)$, $y_0 = y_1 = 0$. Le problème admet pour solution la fonction

$$y_\varepsilon(x) = 20 - de^x + (20 - d)(\operatorname{cth}\gamma \operatorname{sh}\gamma x - \operatorname{ch}\gamma x) +$$
$$+ de(20 - de)\frac{\operatorname{sh}\gamma x}{\operatorname{sh}\gamma}, \text{ où } d = \frac{10}{1-\varepsilon^2}, \gamma = \frac{1}{\varepsilon}.$$

On trouve dans le tableau 6.5 les solutions y_ε calculées par cette formule pour $\varepsilon_1 = 0{,}1$, $\varepsilon_2 = 0{,}05$ et $\varepsilon_0 = 0{,}01$, la solution $u = f/a$ du problème dégénéré et la solution extrapolée Y_{ε_0}.

Tableau 6.5

x	y_{ε_1}	y_{ε_2}	y_{ε_0}	Y_{ε_0}	u
0	0	0	0	0	10
0,01	0,3406	0,7074	6,2201	6,2356	9,8995
0,02	1,5905	3,0860	8,4437	8,4554	9,7980
0,03	2,2582	4,1553	9,1966	9,2033	9,6955
0,04	2,8515	5,0838	9,4077	9,4114	9,5919
0,05	3,3774	5,7914	9,4189	9,4210	9,4873
0,06	3,8421	6,3506	9,3558	9,3573	9,3816
0,07	4,2514	6,7882	9,2647	9,2660	9,2749
0,08	4,6104	7,1261	9,1627	9,1638	9,1671
0,09	4,9238	7,3820	9,0559	9,0570	9,0583
0,10	5,1958	7,5706	8,9468	8,9478	8,9483

On voit que les résultats de l'extrapolation concordent bien avec la solution exacte au voisinage du point $x = 0$, et cela en dépit d'un écart sensible entre la solution cherchée y_{ε_0} et les valeurs y_{ε_1}, y_{ε_2}, u utilisées dans l'extrapolation.

CHAPITRE 7

ANNEXE

Ce chapitre comprend plusieurs résultats simples qu'on a utilisés plus d'une fois dans les chapitres précédents.

7.1. Développement des quotients différentiels suivant le pas du réseau

Les résultats de ce paragraphe sont des conséquences simples de la formule de Taylor.

LEMME 1.1. *Si les points x, $x \pm h$ appartiennent au segment $[0, 1]$, toute fonction $v(x) \in C^m[0, 1]$ admet les développements*

$$v_{\mathring{x}}(x) = \frac{v(x + h/2) - v(x - h/2)}{h} =$$

$$= \sum_{i=0}^{r} \frac{h^{2i}}{(2i+1)!} \frac{v^{(2i+1)}(x)}{2^{2i}} + 2^{-m+1} h^{m-1} \varkappa_1, \quad (1.1)$$

où $r = [m/2] - 1$;

$$v_{\bar{\mathring{x}}}(x) = \frac{v(x + h/2) + v(x - h/2)}{2} = \sum_{i=0}^{q} \frac{h^{2i} v^{(2i)}(x)}{(2i)!\, 2^{2i}} + 2^{-m} h^m \varkappa_2, \quad (1.2)$$

où $q = [(m-1)/2]$;

$$v_x(x) = \frac{v(x + h) - v(x)}{h} = \sum_{i=0}^{m-2} \frac{h^i}{(i+1)!} v^{(i+1)}(x) + h^{m-1} \varkappa_3 \quad (1.3)$$

et

$$v_{\bar{x}}(x) = \frac{v(x) - v(x - h)}{h} = \sum_{i=0}^{m-2} \frac{(-h)^i}{(i+1)!} v^{(i+1)}(x) + h^{m-1} \varkappa_4, \quad (1.4)$$

toutes les quantités $\varkappa_i$ étant indépendantes de x et h et bornées par

$$\frac{1}{m!} \max_{[0,1]} |v^{(m)}|.$$

Ces développements sont à la base du

LEMME 1.2. *On suppose que les points* x, $x \pm h$ *appartiennent au segment* $[0, 1]$ *et que* $p \in C^m[0, 1]$, $v \in C^{m+1}([0, 1])$. *On a le développement*

$$(pv_{\bar{x}})_{\hat{x}}|_x = \frac{p(x+h/2)\left(v(x+h)-v(x)\right) - p(x-h/2)\left(v(x)-v(x-h)\right)}{h^2} =$$

$$= \sum_{s=0}^{q} h^{2s} \sum_{0 \leqslant i+j \leqslant s} \frac{\left(p(x)\, v^{(2i+1)}(x)\right)^{(2j+1)}}{(2i+1)!\,(2j+1)!\,4^{i+j}} + h^{m-1}\varkappa_5, \quad (1.5)$$

où $q = [(m-1)/2]$,

$$|\varkappa_5| \leqslant c_1 \max_{0 \leqslant s \leqslant m} |p^{(s)}| \max_{0 \leqslant s \leqslant m+1} |v^{(s)}|, \quad (1.6)$$

avec la constante c_1 *indépendante de* x, h, p, v.

DÉMONSTRATION. La formule (1.1) entraîne

$$v_{\bar{x}}\left(x - \frac{h}{2}\right) = \frac{v(x)-v(x-h)}{h} =$$

$$= \sum_{k=0}^{q} \frac{h^{2k}}{4^k(2k+1)!} v^{(2k+1)}\left(x - \frac{h}{2}\right) + h^m \varkappa^-,$$

$$v_{\bar{x}}\left(x + \frac{h}{2}\right) = \frac{v(x+h)-v(x)}{h} =$$

$$= \sum_{k=0}^{q} \frac{h^{2k}}{4^k\, 2(k+1)!} v^{(2k+1)}\left(x + \frac{h}{2}\right) + h^m \varkappa^+,$$

où

$$|\varkappa^{\pm}| \leqslant \frac{1}{2^m(m+1)!} \max_{[0,1]} |v^{(m+1)}|. \quad (1.7)$$

On multiplie ces différences divisées par les valeurs correspondantes de p, on effectue la soustraction, on divise par h et on applique le lemme 1.1. à chaque terme $pv^{(2k+1)}$, il vient

$$p(x+h/2)\, v_{\bar{x}}(x+h/2) - p(x-h/2)\, v_{\bar{x}}(x-h/2) =$$

$$= \sum_{k=0}^{q} \frac{h^{2k}}{4^k(2k+1)!} \frac{p(x+h/2)\, v^{(2k+1)}(x+h/2) - p(x-h/2)\, v^{(2k+1)}(x-h/2)}{h} +$$

$$+ h^{m-1}\left(p(x+h/2)\varkappa^+ - p(x-h/2)\varkappa^-\right) = \sum_{k=0}^{q} \frac{h^{2k}}{4^k(2k+1)!} \times$$

$$\times \left[\sum_{j=0}^{q-k} \frac{h^{2j}}{4^j(2j+1)!} \left(p(x)\, v^{(2k+1)}(x)\right)^{(2j+1)} + \mu_k h^{m-2k-1}\right] +$$

$$+ h^{m-1}\left(p(x+h/2)\varkappa^+ - p(x-h/2)\varkappa^-\right), \quad (1.8)$$

où

$$|\mu_k| \leqslant \frac{1}{2^{m+1-2k}(m+1-2k)!} \max_{[0,1]} \left|\left(p v^{(2k+1)}\right)^{(m+1-2k)}\right| \leqslant$$

$$\leqslant \frac{1}{(m+1-2k)!} \max_{\substack{0 \leqslant s \leqslant m \\ [0,1]}} |p^{(s)}| \max_{\substack{0 \leqslant s \leqslant m+1 \\ [0,1]}} |v^{(s)}|. \quad (1.9)$$

La dernière inégalité découle de la formule d'un produit m fois dérivable. En identifiant les termes de même degré en h de (1.8), on aboutit à (1.5), et l'estimation (1.6) résulte de (1.7) et (1.9).

LEMME 1.3. *Soit* $v \in C^{l+\alpha}[x-h, x+h]$, *avec* $l \geqslant 2$ *entier et* $\alpha \in (0, 1)$. *On a l'égalité*

$$v_{\bar{x}\bar{x}}(x) = \frac{v(x+h) - 2v(x) + v(x-h)}{h^2} = 2\sum_{i=0}^{s} h^{2i} \frac{v^{(2i+2)}(x)}{(2i+2)!} +$$
$$+ h^{l-2+\alpha} \rho(x),$$

où $s = [l/2] - 1$,

$$|\rho(x)| \leqslant \frac{2}{l!} \|v\|_{C^{l+\alpha}[x-h,\, x+h]}.$$

DÉMONSTRATION. On a par la formule de Taylor

$$v(x \pm h) = \sum_{i=0}^{l-1} \frac{(\pm h)^i}{i!} v^{(i)}(x) + \frac{(\pm h)^l}{l!} v^{(l)}(\xi^\pm). \quad (1.10)$$

La dérivée $v^{(l)}$ est continue höldérienne d'exposant α, si bien que

$$\frac{|v^{(l)}(\xi^\pm) - v^{(l)}(x)|}{|\xi^\pm - x|^\alpha} \leqslant c_3 = \|v\|_{C^{l+\alpha}[x-h,\, x+h]}.$$

On a par suite de la condition $|\xi^\pm - x| \leqslant h$:

$$|v^{(l)}(\xi^\pm) - v^{(l)}(x)| \leqslant c_3 h^\alpha,$$

auquel cas la formule (1.10) entraîne

$$v(x \pm h) = \sum_{i=0}^{l} \frac{(\pm h)^i}{i!} v^{(i)}(x) + h^{l+\alpha}\rho^\pm(x), \quad (1.11)$$

où $|\rho^\pm(x)| \leqslant \frac{c_3}{l!}$.

On utilise le développement (1.11) pour calculer $v_{\bar{x}\bar{x}}(x)$.

On a

$$v_{\bar{x}x}(x) = 2 \sum_{i=0}^{s} \frac{h^{2i}}{(2i+2)!} v^{(2i+2)}(x) + h^{l-2+\alpha}(\rho^+(x) + \rho^-(x)),$$

d'où

$$c_2 = \frac{2c_3}{l!} = \frac{2}{l!} \|v\|_{C^{l+\alpha}[x-h,\, x+h]},$$

c.q.f.d.

7.2. Sur la résolution des systèmes d'équations spéciaux

Soit le système d'équations en γ_i

$$\begin{aligned} &\sum_{i=1}^{s+1} \gamma_i = 1, \\ &\sum_{i=1}^{s+1} \gamma_i \mu_i^l = 0, \qquad l = 1, \ldots, s. \end{aligned} \tag{2.1}$$

Il est connu (voir [102]) que le déterminant de Vandermonde

$$V(\nu_1, \nu_2, \ldots, \nu_{s+1}) = \det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \nu_1 & \nu_2 & \cdots & \nu_{s+1} \\ \nu_1^2 & \nu_2^2 & \cdots & \nu_{s+1}^2 \\ \cdot & \cdot & \cdots & \cdot \\ \nu_1^s & \nu_2^s & \cdots & \nu_{s+1}^s \end{bmatrix}$$

est calculé par la formule

$$V(\nu_1, \nu_2, \ldots, \nu_{s+1}) = \prod_{i<j} (\nu_j - \nu_i). \tag{2.2}$$

Le système (2.1) est donc compatible si et seulement si tous les ν_i sont distincts deux à deux. Le problème (2.1) est alors abordé par la méthode de Cramer:

$$\gamma_i = \frac{V(\mu_1, \ldots, \mu_{i-1}, 0, \mu_{i+1}, \ldots, \mu_{s+1})}{V(\mu_1, \ldots, \mu_{i-1}, \mu_i, \mu_{i+1}, \ldots, \mu_{s+1})}, \tag{2.3}$$

où l'on pose $0^0 = 1$. On a, en particulier, le

LEMME 2.1. *Si le système* (2.1) *a ses coefficients* $\mu_i = \dfrac{1}{i}$, *alors*

$$\gamma_k = \frac{(-1)^{s-k+1} k^{s+1}}{k!\,(s-k+1)!}, \qquad k = 1, \ldots, s+1.$$

DÉMONSTRATION. On effectue dans (2.3) les substitutions par la formule (2.2) et on réduit les termes identiques, il vient

$$\gamma_k = \prod_{i=1}^{k-1} \frac{-\mu_i}{\mu_k - \mu_i} \prod_{i=k+1}^{s+1} \frac{\mu_i}{\mu_i - \mu_k}. \tag{2.4}$$

On remplace μ_i par l/i, et on a l'affirmation voulue.

On démontre de même le

LEMME 2.2. *Si le système* (2.1) *a ses coefficients* $\mu_i = 1/i^2$, *alors*

$$\gamma_k = 2 \cdot \frac{(-1)^{s-k+1}\, k^{2s+2}}{(s+k+1)!\,(s-k+1)!}, \qquad k = 1, \ldots, s+1.$$

LEMME 2.3. *Si les coefficients du système d'équations* (2.1) *remplissent la condition*

$$\mu_i/\mu_{i+1} \geqslant 1 + c, \quad i = 1, \ldots, s,$$

avec c une constante positive, alors la solution du système est évaluée par

$$|\gamma_i| \leqslant \left(\frac{1+c}{c}\right)^s, \qquad i = 1, \ldots, s+1.$$

DÉMONSTRATION. On transforme la formule (2.4):

$$|\gamma_k| = \prod_{i=1}^{k-1} \frac{1}{\mu_i/\mu_k - 1} \prod_{i=k+1}^{s+1} \left(1 + \frac{1}{\mu_k/\mu_i - 1}\right)$$

et on récrit la condition du lemme:

$$\mu_i/\mu_k \geqslant (1+c)^{k-i} \geqslant 1 + c \text{ pour } i < k.$$

Les estimations des facteurs de l'égalité précédente fournissent

$$|\gamma_k| \leqslant \left(\frac{1}{c}\right)^{k-1} \cdot \left(1 + \frac{1}{c}\right)^{s-k+1} = \frac{(1+c)^{s-k+1}}{c^s} \leqslant \left(\frac{1+c}{c}\right)^s,$$

ce qui démontre le lemme.

LEMME 2.4. *Si le système d'équations*

$$\sum_{i=1}^{s+1} \beta_i = 1,$$

$$\sum_{i=1}^{s+1} \beta_i \nu_i^{2l} = 0, \qquad l = 1, \ldots, s,$$

a des coefficients tels que

$$\frac{\nu_i}{\nu_{i+1}} \geqslant 1 + d, \quad i = 1, \ldots, s; \quad d > 0, \tag{2.5}$$

alors le système admet une solution unique vérifiant l'inégalité

$$|\beta_k| \leqslant \left(\frac{1 + 2d + d^2}{2d + d^2}\right)^s, \quad k = 1, \ldots, s.$$

La démonstration s'inspire du lemme 2.3 à condition de poser $\nu_i^2 = \mu_i$. L'inégalité (2.5) devient

$$\frac{\mu_i}{\mu_{i+1}} \geqslant (1 + d)^2 = 1 + 2d + d^2,$$

et $c = 2d + d^2$ dans la condition du lemme 2.3.

Soit le système d'équations en α_i

$$\sum_{i=1}^{s} \alpha_i \mu_l^{i-1} = \mu_l^s, \qquad l = 1, \ldots, s + 1. \tag{2.6}$$

Le déterminant de la matrice du système est égal au déterminant de Vandermonde $V(\mu_1, \ldots, \mu_{s+1})$ dont il est le transposé (voir [102]). Si μ_i sont distincts deux à deux, (2.6) possède donc une solution unique. On demande le coefficient

$$\alpha_1 = \frac{\det \begin{bmatrix} \mu_1^{s+1} & \mu_1 & \cdots & \mu_1^s \\ \mu_2^{s+1} & \mu_2 & \cdots & \mu_2^s \\ \mu_{s+1}^{s+1} & \mu_{s+1} & \cdots & \mu_{s+1}^s \end{bmatrix}}{V(\mu_1, \mu_2, \ldots, \mu_{s+1})}.$$

On transforme le numérateur: dans chaque ligne du déterminant on met μ_1 en facteur, puis on met la première colonne à la dernière place, il vient

$$\alpha_1 = \frac{(-1)^s \mu_1\mu_2 \ldots \mu_{s+1} V(\mu_1, \mu_2, \ldots, \mu_{s+1})}{V(\mu_1, \mu_2, \ldots, \mu_{s+1})} = (-1)^s \mu_1 \ldots \mu_{s+1}. \tag{2.7}$$

LEMME 2.5. *On admet que μ_i sont pour tous les $i = 1, \ldots, s$ distincts deux à deux et que γ_i sont obtenus moyennant le système (2.1). On a*

$$\sum_{i=1}^{s+1} \gamma_i \mu_i^{s+1} = (-1)^s \mu_1\mu_2 \cdots \mu_{s+1}.$$

DÉMONSTRATION. On transforme le premier membre par les relations (2.6) :

$$\sum_{i=1}^{s+1} \gamma_i \mu_i^{s+1} = \sum_{i=1}^{s+1} \gamma_i \sum_{j=1}^{s+1} \alpha_j \mu_i^{j-1},$$

on change l'ordre de sommation

$$\sum_{i=1}^{s+1} \gamma_i \mu_i^{s+1} = \sum_{j=1}^{s+1} \alpha_j \sum_{i=1}^{s+1} \gamma_i \mu_i^{j-1},$$

et on utilise (2.1), il vient

$$\sum_{i=1}^{s+1} \gamma_i \mu_i^{s+1} = \alpha_1 = (-1)^s \mu_1 \mu_2 \dots \mu_{s+1},$$

c.q.f.d.

S'agissant des puissances paires, on énonce un résultat analogue (sans le démontrer car il découle du lemme 2.5) sous forme de

LEMME 2.6. *Si l'on est dans les hypothèses du lemme* 2.4, *alors*

$$\sum_{i=1}^{s+1} \beta_i \nu_i^{2s+2} = (-1)^s \nu_1^2 \nu_2^2 \dots \nu_{s+1}^2.$$

LEMME 2.7. *Soit* $\varepsilon_i \neq 0$ *pour tout* $i = 1, \dots, k$. *On a*

$$W = \det \begin{bmatrix} \varepsilon_1^{-1} & \varepsilon_2^{-1} & \dots & \varepsilon_k^{-1} \\ \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_k \\ \varepsilon_1^2 & \varepsilon_2^2 & \dots & \varepsilon_k^2 \\ \cdot & \cdot & \cdot & \cdot \\ \varepsilon_1^{k-1} & \varepsilon_2^{k-1} & \dots & \varepsilon_k^{k-1} \end{bmatrix} = \sum_{i=1}^{k} \varepsilon_i^{-1} V(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k).$$

DÉMONSTRATION. On multiplie chaque i-ième colonne par ε_i, auquel cas

$$\varepsilon_1 \varepsilon_2 \dots \varepsilon_k W = \det \begin{bmatrix} 1 & 1 & \dots & 1 \\ \varepsilon_1^2 & \varepsilon_2^2 & \dots & \varepsilon_k^2 \\ \varepsilon_1^3 & \varepsilon_2^3 & \dots & \varepsilon_k^3 \\ \cdot & \cdot & \cdot & \cdot \\ \varepsilon_1^k & \varepsilon_2^k & \dots & \varepsilon_k^k \end{bmatrix}. \tag{2.8}$$

Soit le déterminant

$$V(t, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_k) = \det \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ t & \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_k \\ t^2 & \varepsilon_1^2 & \varepsilon_2^2 & \dots & \varepsilon_k^2 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ t^k & \varepsilon_1^k & \varepsilon_2^k & \dots & \varepsilon_k^k \end{bmatrix}. \tag{2.9}$$

On le développe suivant la première colonne et on constate que le déterminant (2.8) est égal au coefficient changé de signe de t de ce développement. Quant à (2.9) tout entier, on a

$$\left\{\prod_{k\geq j>i\geq 1}(\varepsilon_j-\varepsilon_i)\right\}\prod_{l=1}^{k}(\varepsilon_l-t)=V(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k)(-1)^k\prod_{l=1}^{k}(t-\varepsilon_l).$$

Le coefficient de t du polynôme $\prod_{l=1}^{k}(t-\varepsilon_l)$ vaut

$$\sum_{l=1}^{k}\frac{\varepsilon_1\ldots\varepsilon_k}{\varepsilon_l}(-1)^{k+1},$$

ce qui donne pour le coefficient de t du polynôme (2.9):

$$-V(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k)\,\varepsilon_1\ldots\varepsilon_k\sum_{l=1}^{k}\varepsilon_l^{-1}.$$

On l'égale à $-\varepsilon_1\ldots\varepsilon_k W$, et on a l'affirmation du lemme.

Donnons un résultat relatif à l'algorithme de Neville [94].

LEMME 2.8. *On suppose que $T_j^{(0)}$ est un jeu donné de nombres pour $j=1,\ldots,s+1$ et que la suite $T_j^{(i)}$ est définie par l'algorithme de Neville*

$$T_j^{(i)}=\frac{\mu_{i+j}\,T_j^{(i-1)}-\mu_j\,T_{j+1}^{(i-1)}}{\mu_{i+j}-\mu_j}, \tag{2.10}$$

$$j=1,\ldots,s-i+1;\; i=1,2,\ldots,s.$$

On a l'égalité

$$T_1^{(s)}=\sum_{k=1}^{s+1}\gamma_k\,T_k^{(0)}. \tag{2.11}$$

avec γ_k solution du système (2.1).

DÉMONSTRATION. Soit $s=1$. Le système (2.1) donne

$$\gamma_1=\frac{-\mu_2}{\mu_1-\mu_2},\qquad \gamma_2=\frac{\mu_1}{\mu_1-\mu_2},$$

si bien que les règles (2.10) et (2.11) coïncident. En effet,

$$T_1^{(1)}=\gamma_2\,T_2^{(0)}+\gamma_1\,T_1^{(0)}.$$

On suppose que le résultat voulu est démontré pour $s-1$, et on en veut la preuve pour s. L'égalité (2.11) étant satisfaite pour $s-1$ implique deux relations

$$T_1^{(s-1)}=\sum_{k=1}^{s}\beta_k\,T_k^{(0)}\quad\text{et}\quad T_2^{(s-1)}=\sum_{k=1}^{s}\delta_k\,T_{k+1}^{(0)}.$$

où les coefficients β_k et δ_k sont solutions des systèmes respectifs

$$\begin{aligned} &\sum_{k=1}^{s} \beta_k = 1, \qquad \sum_{k=1}^{s} \mu_k^l \beta_k = 0, \qquad l = 1, 2, \dots, s-1; \\ &\sum_{k=1}^{s} \delta_k = 1, \qquad \sum_{k=1}^{s} \mu_{k+1}^l \delta_s = 0, \qquad l = 1, 2, \dots, s-1. \end{aligned} \tag{2.12}$$

On calcule $T_1^{(s)}$:

$$T_1^{(s)} = \frac{\mu_{s+1} T_1^{(s-1)} - \mu_1 T_2^{(s-1)}}{\mu_{s+1} - \mu_1} = \frac{\mu_{s+1} \beta_1}{\mu_{s+1} - \mu_1} T_1^{(0)} +$$

$$+ \sum_{k=2}^{s} \frac{\mu_{s+1} \beta_k - \mu_1 \delta_{k-1}}{\mu_{s+1} - \mu_1} T_k^{(0)} - \frac{\mu_1 \delta_s}{\mu_{s+1} - \mu_1} T_{s+1}^{(0)} = \sum_{k=1}^{s+1} \rho_k T_k^{(0)}. \tag{2.13}$$

où

$$\rho_1 = \frac{\mu_{s+1} \beta_1}{\mu_{s+1} - \mu_1}, \qquad \rho_{s+1} = \frac{\mu_1 \delta_s}{\mu_{s+1} - \mu_1},$$

$$\rho_k = \frac{\mu_{s+1} \beta_k - \mu_1 \delta_{k-1}}{\mu_{s+1} - \mu_1}, \qquad k = 2, 3, \dots, s.$$

On montre que ces coefficients satisfont au système (2.1.) On a compte tenu de (2.12):

$$\sum_{k=1}^{s+1} \rho_k = \frac{\mu_{s+1}}{\mu_{s+1} - \mu_1} \sum_{k=1}^{s} \beta_k - \frac{\mu_1}{\mu_{s+1} - \mu_1} \sum_{k=1}^{s} \delta_k - \frac{\mu_{s+1} - \mu_1}{\mu_{s+1} - \mu_1} = 1;$$

$$\sum_{k=1}^{s+1} \mu_k^l \rho_k = \Big[\mu_{s+1} \beta_1 \mu_1^l + \sum_{k=2}^{s} (\mu_{s+1} \beta_k - \mu_1 \delta_{k-1}) \mu_k^l -$$

$$- \mu_1 \delta_s \mu_1^l \Big] (\mu_{s+1} - \mu_1)^{-1} =$$

$$= \Big[\mu_{s+1} \sum_{k=1}^{s} \beta_k \mu_k^l - \mu_1 \sum_{k=1}^{s} \delta_k \mu_{k+1}^l \Big] (\mu_{s+1} - \mu_1)^{-1}. \tag{2.14}$$

Si $l = 1, 2, \dots, s-1$, les deux sommes entre crochets s'annulent par suite de (2.12). Si $l = s$, alors on a, conformément au lemme 2.5,

$$\sum_{k=1}^{s} \beta_k \mu_k^l = (-1)^{s-1} \mu_1 \mu_2 \dots \mu_s,$$

$$\sum_{k=1}^{s} \delta_k \mu_{k+1}^l = (-1)^{s-1} \mu_2 \mu_3 \dots \mu_{s+1}.$$

L'expression (2.14) fournit maintenant

$$\sum_{k=1}^{s+1} \mu_k^s \rho_k = 0.$$

Ainsi, ρ_k vérifie le système (2.1). Comme il y a unicité pour μ_i distincts deux à deux, on a $\rho_k = \gamma_k$, et (2.11) et (2.13) ont même second membre.

7.3. Sur les polynômes d'interpolation de Lagrange

Soit $f(x)$ une fonction continue sur le segment [0, 1] dont on connaît les valeurs en s points régulièrement espacés $x_i = x_1 + (i-1)h$ de [0, 1]. La fonction

$$L_s(x) = \sum_{i=1}^{s} f(x_i) \prod_{\substack{j \neq i \\ j=1}}^{s} \frac{x - x_j}{x_i - x_j} \tag{3.1}$$

est par définition le polynôme d'interpolation de Lagrange. Si $f \in C^s [0, 1]$, on a pour le point $x \in [x_1, x_s]$ (voir [67]):

$$f(x) - L_s(x) = \frac{1}{s!} f^{(s)}(\xi)\, \omega_s(x), \tag{3.2}$$

où $\xi \in [0, 1]$ et

$$\omega_s(x) = \prod_{j=1}^{s} (x - x_j). \tag{3.3}$$

On se propose d'évaluer ces quantités en fonction des puissances de h. S'agissant des points intérieurs à $[x_1, x_2]$, le maximum de $\omega_2(x)$ en valeur absolue est atteint au milieu du segment, et il est égal à $h^2/4$. Aussi

$$|\omega_2(x)| \leqslant h^2/4, \qquad x \in [x_1, x_2].$$

On démontre la majoration

$$|\omega_l(x)| \leqslant \frac{h^l}{4}(l-1)!, \quad x \in [x_1, x_l]. \tag{3.4}$$

L'inégalité est évidente pour $l = 2$. On la suppose juste pour l et on la démontre pour $l+1$. Deux cas peuvent se présenter. Si $x \in [x_1, x_l]$, alors

$$|\omega_{l+1}(x)| \leqslant |\omega_l(x)|\,|x - x_{l+1}| \leqslant \frac{h^l}{4}(l-1)!\, lh = \frac{h^{l+1}}{4} l!,$$

ce qui établit (3.4). Si $x \in [x_l, x_{l+1}]$, on a les inégalités évidentes

$$|x - x_l| \, |x - x_{l+1}| \leq \frac{h^2}{4}, \, |x - x_i| \leq (l - i + 1) h.$$

Aussi

$$|\omega_{l+1}(x)| \leq h^{l-1} \, l(l-1) \ldots 2 \frac{h^2}{4} = \frac{h^{l+1}}{4} l\,!.$$

On réunit les relations (3.2) à (3.4):

$$|f(x) - L_s(x)| \leq \frac{h^s}{4s} \|f\|_{C^s[0,1]}, \; x \in [x_1, x_s]. \tag{3.5}$$

Si $f(x)$ n'est pas suffisamment régulière, l'erreur est de l'ordre de grandeur inférieur.

Lemme 3.1. *Soit* $s > k > 0$ *et* $f \in C^k[0, 1]$. *On a*

$$|f(x) - L_s(x)| \leq \frac{h^k}{k 2^{k-s+2}} \|f\|_{C^k[0,1]}, \; x \in [x_1, x_s]. \tag{3.6}$$

Démonstration. On utilise [67; pp. 37-38] et on désigne par $f(x_1, \ldots, x_l)$ la différence divisée d'ordre l. On écrit alors $L_s(x)$ sous forme de suite de polynômes d'interpolation

$$L_s(x) = L_k(x) + \big(L_{k+1}(x) - L_k(x)\big) + \ldots + \big(L_s(x) - L_{s-1}(x)\big).$$

Comme

$$|L_{j+1}(x) - L_j(x)| = f(x_1, \ldots, x_j) \, \omega_j(x),$$

on a

$$f(x) - L_s(x)| \leq |f(x) - L_k(x)| + |L_{k+1}(x) - L_k(x)| + \ldots$$
$$\ldots + |L_s(x) - L_{s-1}(x)| \leq \frac{h^k}{4k} \|f\|_{C^k[0,1]} +$$
$$+ |f(x_1, \ldots, x_{k+1}) \, \omega_k(x)| + \ldots + |f(x_1 \ldots, x_s) \, \omega_{s-1}(x)|. \tag{3.7}$$

On démontre par récurrence qu'on a pour $f \in C^k [0, 1]$ quelconque et tout $l = k + 1, \ldots, s$ l'inégalité

$$|f(x_1, x_2, \ldots, x_l)| \leq \frac{h^{k+1-l}}{2^{k+1-l}(l-1)\,!} \|f\|_{C^k[x_1, x_l]}. \tag{3.8}$$

Soit $l = k + 1$. On établit dans [67] que

$$f(x_1, x_2, \ldots, x_{k+1}) = \frac{f(0)}{k\,!}, \text{ où } 0 \in [x_1, x_{k+1}],$$

si bien que l'estimation (3.8) a manifestement lieu. On suppose (3.8) juste pour un entier $l > k + 1$. Dans ce cas,

$$|f(x_2, x_3, \ldots, x_{l+1})| \leqslant \frac{h^{k+1-l}}{2^{k+1-l}(l-1)!} \| f \|_{C^k[x_1, x_{l+1}]}$$

découle de (3.8) pour $g(x) = f(x+h)$:

$$|f(x_2, x_3, \ldots, x_{l+1})| = |g(x_1, \ldots, x_l)| \leqslant \frac{h^{k+1-l}}{2^{k+1-l}(l-1)!} \| g \|_{C^k[x_1, x_{l+1}]}.$$

Par définition d'une différence divisée

$$f(x_1, \ldots, x_{l+1}) = \frac{f(x_2, \ldots, x_{l+1}) - f(x_1, \ldots, x_l)}{x_{l+1} - x_1},$$

d'où

$$|f(x_1, \ldots, x_{l+1})| \leqslant \frac{h^{k+1-l}}{2^{k+1-l}(l-1)!} \, \frac{\|f\|_{C^k[x_2, x_{l+1}]} + \|f\|_{C^k[x_1, x_l]}}{lh} \leqslant$$

$$\leqslant \frac{h^{k-l}}{2^{k-l}\, l!} \| f \|_{C^k[x_1, x_{l+1}]}.$$

i.e. l'estimation (3.8) est juste.

On utilise (3.8) pour prolonger la chaîne d'estimations (3.7):

$$|f(x) - L_s(x)| \leqslant \frac{h^k}{4k} \| f \|_{C^k[0,1]} +$$

$$+ \frac{h^k}{4k} \| f \|_{C^k[0,1]} + \ldots + \frac{2^{s-k-1}\, h^{k+1-s}}{4(s-1)} \| f \|_{C^k[0,1]} \leqslant$$

$$\leqslant \frac{h^k}{4} \| f \|_{C^k[0,1]} \left(\frac{2}{k} + \frac{2}{k+1} + \ldots + \frac{2^{s-k-1}}{s-1} \right).$$

Etant donné $s > k \geqslant 1$, on a

$$|f(x) - L_s(x)| \leqslant \frac{h^k}{4} \| f \|_{C^k[0,1]} \frac{2^{s-k}}{k},$$

et le lemme 3.1 se trouve démontré.

On veut évaluer les polynômes

$$\alpha_i(x) = \prod_{\substack{j \neq i \\ j=1}}^{s} \frac{x - x_j}{x_i - x_j} \tag{3.9}$$

caractéristiques de la grandeur de l'erreur dans la formule d'interpolation si les valeurs données $f(x_i)$ sont affectées d'erreurs.

LEMME 3.2. *On a pour les points* $x \in [x_1, x_s]$

$$\sum_{i=1}^{s} |\alpha_i(x)| \leqslant 2^{s-1}. \tag{3.10}$$

DÉMONSTRATION. S'agissant des points équidistants (voir [67]), on a la formule

$$|\alpha_i| = C_{s-1}^{i-1} \frac{t(t-1)\ldots(t-i+2)(t-i)\ \ldots(t-s+1)}{(s-1)!},$$

où $t = (x - x_1)/h$, donc $t \in [0, s-1]$. On vérifie aisément que le numérateur est au plus égal à la constante $(s-1)!$. Alors

$$\sum_{i=1}^{s} |\alpha_i| \leqslant \sum_{i=1}^{s} C_{s-1}^{i-1} = 2^{s-1}.$$

BIBLIOGRAPHIE

[1] Агошков В. И., *О вариационной форме интегрального тождества Г. И. Марчука.* ВЦ СО АН СССР, Новосибирск, 1977.

[2] Алибеков Х. А., Соболевский П. Е., Об устойчивости разностных схем для параболических уравнений, *ДАН СССР*, 1977, 232, nº 4, с. 737-740.

[3] Артемьев С. С., Демидов Г. В., *А*-устойчивый метод типа Розенброка четвертого порядка точности решения задачи Коши для жестких систем обыкновенных дифференциальных уравнений. In: *Некоторые проблемы вычислительной и прикладной математики*, Новосибирск, Наука, 1975, с. 214-219.

[4] Валиуллин А. Н., *Разностные схемы повышенной точности для задач математической физики*, НГУ, Новосибирск, 1970.

[5] Вишик М. И., Люстерник Л. А., Регулярное вырождение и пограничный слой для линейных дифференциальных уравнений с малым параметром, *УМН*, 1957, 12, вып. 5 (77), с. 3-122.

[6] Волков Е. А., Об одном способе повышения точности метода сеток, *ДАН СССР*, 1954, 96, nº 4, с. 685-688.

[7] Волков Е. А., Исследование одного способа повышения точности метода сеток при решении уравнения Пуассона. In: *Вычислительная математика*, М., 1957, nº 1, с. 62-80.

[8] Волков Е. А., Дифференциальные свойства решений краевых задач для уравнения Лапласа и Пуассона на прямоугольнике, *Труды МИАН СССР*, 1965, 77, с. 89-112.

[9] Волков Е. А., Решение задачи Дирихле методом уточнений разностями высших порядков, ч. I, *Дифф. уравнения*, 1965, 1, nº 7, с. 946-960.

[10] Волков Е. А., Решение задачи Дирихле методом уточнений разностями высших порядков, ч. II, *Дифф. уравнения*, 1965, 1, nº 8, с. 1070-1084.

[11] Волков Е. А., Приближенное решение уравнений Лапласа и Пуассона в весовых пространствах Гёльдера, *Труды МИАН СССР*, 1972, 128, с. 76-112.

[12] Волков Е. А., О методе регулярных составных сеток для уравнения Лапласа на многоугольниках, *Труды МИАН СССР*, 1976, 140, с. 68-102

[13] Давиденко Д. Ф., Об одном методе построения разностных уравнений при решении методом сеток внутренней задачи Дирихле для уравнения Пуассона, *Укр. матем. журнал*, 1961, 13, nº 4, с. 92-96.

[14] Демидов Г. В., Об одном методе построения устойчивых схем высокого порядка аппроксимации, *Числ. методы механики сплошной среды*, 1970, 1, nº 6, с. 60-69.

[15] Иванов В. К., О некорректно поставленных задачах, Матем. сб., 1963, 61, nº 2, с. 111-113.

[16] Ильин В. П., *Разностные методы решения эллиптических уравненний*, НГУ, Новосибирск, 1970.

[17] Ильин В. П., *Численные методы решения задач электрооптики*, Новосибирск, Наука, 1974.

[18] Кондратьев В. А., Краевые задачи для эллиптических уравнений в областях с коническими или угловыми точками, *Труды ММО*, 1967, 16, с. 109-192.

[19] Кочергин В. П., Климок В. И., Щербаков А. В., О вычислении градиентов разностного решения, *Числ. методы механики сплошной среды*, 1977, 8, nº 3.

[20] Кочергин В. П., Щербаков А. В., Исследование разностных схем для эллиптического уравнения с малым параметром при старших производных. In: *Численные модели океанических циркуляций*, Новосибирск, ВЦ СО АН СССР, с. 7-24.

[21] Крейн С. Г., *Линейные дифференциальные уравнения в банаховом пространстве*, М., Наука, 1967.

[22] Крылов В. И., Бобков В. В., Монастырный П. И., *Вычислительные методы*, Минск, Вышейша школа, 1975, Т. 2.

[23] Крылов В. И., Бобков В. В., Монастырный П. И., *Вычислительные методы*, М., Наука, 1976, Т. 1; 1977, Т. 2.

[24] Кузнецов Ю. А., Шайдуров В. В., О равномерной сходимости разностных схем, I. In: *Вычислительные методы линейной алгебры*, Новосибирск, ВЦ СО АН СССР, 1972, с. 70-92.

[25] Ладыженская О. А., Солонников В. А., Уральцева Н. Н., *Линейные и квазилинейные уравнения параболического типа*, М., Наука, 1967.

[26] Ладыженская О. А., Уральцева Н. Н., *Линейные и квазилинейные уравнения эллиптического типа*, М., Наука, 1973.

[27] Лебедев В. И., О задаче Дирихле и Неймана на треугольных и шестиугольных сетках, *ДАН СССР*, 1961, 138, nº 1, с. 33-36.

[28] Лебедев В. И., О четырехточечных схемах повышенной точности, *ДАН СССР*, 1962, 142, nº 3, с. 526-529.

[29] Марчук Г. И., *Методы расчета ядерных реакторов*, М., Госатомиздат, 1961.

[30] Марчук Г. И., *Методы вычислительной математики*, Новосибирск, Наука, 1973.

[31] Марчук Г. И., Лебедев В. И., *Численные методы в теории переноса нейтронов*, М., Атомиздат, 1971.

[32] Микеладзе Ш. Е., О численном интегрировании уравнений эллиптического и параболического типа, *Изв. АН СССР*. Сер. матем., 1941, 5, nº 1, с. 57-73.

[33] Михлин С. Г., *Вариационные методы в математической физике*, М., Наука, 1970.

[34] Михлин С. Г., *Линейные уравнения в частных производных*, М., Высшая школа, 1977.

[35] Морозов В. А., О псевдорешениях, *ЖВМ и МФ*, 1969, 9, nº 6, с. 1387-1391.

[36] Морозов В. А., Метод регуляризации и решение систем линейных алгебраических уравнений. In: *Вычислительные методы линейной алгебры*, Новосибирск, ВЦ СО АН СССР, 1972, с. 64-69.

[37] Оганесян Л. А., Ривкинд В. Я., Руховец Л. А., Вариационно-разностные методы решения эллиптических уравнений, ч. I, *Дифференциальные уравнения и их применение*, вып. 5, Вильнюс, 1973.

[38] Оганесян Л. А., Ривкинд В. Я., Руховец Л. А., Вариационно-разностные методы решения эллиптических уравнений, ч. II, *Дифференциальные уравнения и их применение*, вып. 8, Вильнюс, 1974.

[39] Оганесян Л. А., Руховец Л. А., Исследование скорости сходимости вариационно-разностных схем для эллиптических уравнений второго порядка в двумерной области с гладкой границей, *ЖВМ и МФ*, 1969, 9, nº 5, с. 1102-1120.

[40] Приказчиков В. Г., Однородные разностные схемы высокого порядка точности для задачи Штурма — Лиувилля, *ЖВМ и МФ*, 1969, 9, nº 2, с. 315-336.

[41] Самарский А. А., Схемы повышенного порядка точности для многомерного уравнения теплопроводности, *ЖВМ и МФ*, 1963, 3, nº 3, с. 431-466.

[42] Самарский А. А., *Введение в теорию разностных схем*, М., Наука, 1971.

[43] Самарский А. А., *Теория разностных схем*, М., Наука, 1977.

[44] Султанова И. А., Эффективные оценки погрешности метода сеток решения краевых задач для уравнений Лапласа и Пуассона на прямоугольнике и специальных треугольниках, *ЖВМ и МФ*, 1971, 11, nº 5, с. 1205-1218.

[45] Тихонов А. Н., О некорректных задачах линейной алгебры и устойчивом методе их решения, *ДАН СССР*, 1965, 163, nº 3, с. 591-594.

[46] Урванцев А. Л., Шайдуров В. В., Метод уточнения для одномерного квазилинейного уравнения диффузии. In: *Вычислительная математика и программирование*, Новосибирск, ВЦ СО АН СССР, 1974, с. 81-90.

[47] Федорова О. А., Вариационно-разностная схема для одномерного уравнения диффузии, *Матем. заметки*, 1975, 17, nº 6, с. 893-898.

[48] Фихтенгольц Г. М., *Курс дифференциального и интегрального исчисления*, М., Наука, 1969, Т. 2.

[49] Фихтенгольц Г. М., *Курс дифференциального и интегрального исчисления*, М., Наука, 1969, Т. 3.

[50] Фрязинов И. В., Экономичные схемы повышенного порядка точности для решения многомерного уравнения параболического типа, *ЖВМ и МФ*, 1969, 9, nº 6, с. 1316-1326.

[51] Фуфаев В. В., К задаче Дирихле для областей с углами, *ДАН СССР*, 1960, 131, nº 1, с. 37-39.

[52] Шайдуров В. В., Об одном методе повышения точности разностных решений, *Числ. методы механики сплошной среды*, 1972, 3, nº 2, с. 96-104.

[53] Шайдуров В. В., Продолжение по параметру в методе регуляризации In: *Вычислительные методы линейной алгебры*, Новосибирск, ВЦ СО АН СССР, 1972, с. 77-85.

[54] Шайдуров В. В., Регуляризация систем с симметричной матрицей. In: *Вычислительная математика и программирование*. Новосибирск, ВЦ СО АН СССР, 1974, с. 91-98.

[55] Шайдуров В. В., *Методы повышения точности приближенных задач*, Новосибирск, НГУ, 1978.

[56] Щербаков А. В., Кочергин В. П., Метод вложенных сеток в задачах динамики океана, *Числ. методы механики сплошной среды*, 1977, 8, nº 2.

[57] Albrecht I., Uhlmann W., Differenzenverfahren für die 1. Randwertaufgabe mit krummlinigen Rändern bei $\Delta u(x, y) = r(x, y u)$, *Z. angew. Math. und Mech.*, 1957, 37, nº 5/6, S. 212-224.

[58] Atkinson F. V., *Discrete and continuous boundary problems*, Academic Press, New York-London, 1964.

[59] Aubin J. P., Behaviour of the error of the approximate solutions of boundary-value problems for linear elliptic operators by Galerkin's and finite-difference methods, *Ann. Sci. Norm. Pisa*, 21, 1961, p. 599-637.

[60] BABUŠKA I., Finite element method for domains with corners, *Computing*, 1970, **6**, nº 3, p. 264-273.

[61] BABUŠKA I., The finite element method with Lagrangian multipliers, *Numer. Math.*, 1973, **21**, nº 16, p. 322-333.

[62] BABUŠKA I., AZIZ A. K., On the angle condition in the finite element method, *SIAM J. on Numer. Anal.*, 1976, **13**. nº 2, p. 214-226.

[63] BABUŠKA I., RHEINBOLDT W., MESZTENYI C., *Self-adaptive refinements in the finite element method.* Technical Report of University of Maryland, 1975.

[64] BABUŠKA I., ROZENZWEIG M. B., A finite element scheme for domains with corners, *Numer, Math.*, 1972, **20**, nº 1, p. 1-21.

[65] BABUŠKA I., VITÁSEK E., PRÁGER M., *Numerical processus in differential Equations*, Intersc. Publ., 1966.

[66] BAKER C. T. H., The deferred approach to the limit for eigenvalues of integral equations, *SIAM J. on Numer. Anal.*, 1971, **8**, nº 1, p. 1-10.

[67] BAKHVALOV N., *Méthodes numériques*, Ed. de Moscou, 1976.

[68] BANK R. E., ROSE D. J., Extrapolated fast direct algorithms for elliptic boundary value problems. In: *Algorithms and complexity. New directions and recent results*, New York, Academic Press, 1976, p. 201-247.

[69] BICKLEY W. G., MICHAELSON S., OSBORNE M. R., Extrapolation on a rectangle, *Proc. Roy Soc.*, 1961, ser. A, nº 262, p. 219.

[70] BRAMBLE J. H., HUBBARD B. E., On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation, *Numer. Math.*, 1962, **4**, nº 4, p. 313-327.

[71] BRAMBLE J. H., SCHATZ A. H., Rayleigh — Ritz — Galerkin methods for Dirichlet's problem using subspaces without boundary conditions, *Comm. Pure Appl. Math.*, 1970, nº 23, p. 653-675.

[72] BREZINSKI C., Conditions d'application et de convergence de procédés d'extrapolation, *Numer. Math.*, 1972, **20**, nº 1, p. 64-79.

[73] BREZINSKI C., Etudes sur les ε- et ρ-algorithmes, *Numer. Math.*, 1971, **17**, nº 2, p. 153-162.

[74] BULIRSCH R., Bemerkungen zur Romberg-Integration, *Numer. Math.*, 1964, **6**, nº 1, p. 6-16.

[75] BULIRSCH R., STOER J., Numerical treatment of ordinary differential equations by extrapolation methods, *Numer. Math.*, 1966, **8**, nº 1, p. 1-13.

[76] BULIRSCH R., STOER J., Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus, *Numer. Math.*, 1964, **6**, nº 5, S. 413-427.

[77] CARTAN H., *Calcul différentiel. Formes différentielles*, Paris, 1967.

[78] COURANT R., HILBERT D., *Methods of mathematical physics*, vol. 1, Intersc. Publ., 1953.

[79] DAHLQUIST G., A special stability problem for linear multistep methods, *BIT*, 1963, nº 3, p. 27-43.

[80] DAHLQUIST G., LINDBERG B., On some implicit one-step methods for stiff differential equations, Stockholm, Royal Inst. of Tech., Dept. of Inf. Proc., TRITA-Na-7302.

[81] DESCLOUX J., *Méthode des éléments finis*, Département de mathématiques, Lausanne, 1973.

[82] ENRIGHT W. H., HULL T. E., Test results on initial value methods for non-stiff ordinary differential equations, *SIAM J. on Numer. Anal.*, 1976, **13**, nº 6, p. 944-961.

[83] FADDEEV D. K., FADDEEVA V. N., *Computational methods of linear algebra*, Freeman, San Francisco, 1963.

[84] Falk R. S., King J. T., A penalty and extrapolation method for the stationary Stokes equations, *SIAM J. on Numer. Anal.*, 1976, **13**, nº 5, p. 814-829.

[85] Forsythe G. E., Wasow W. R., *Finite Difference Methods for partial differential equations*, New York - London, 1960.

[86] Fox L., Accuracy and precision of methods. In: *Numerical Solution of Ordinary and Partial Differential Equations*, Oxford, Pergamon Press, 1967, p. 106-111, 205-312.

[87] Fox L., Mayers D. F., *Computing Methods for Scientists and Enginers*, Oxford, Oxford University Press, 1968.

[88] Gerschgorin S. A., Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen, *Z. angew. Math. und Mech.*, 1930, **10**, S. 373-382.

[89] Godounov S., *Equations de la physique mathématique*, Ed. de Moscou, 1973.

[90] Godounov, S. K., Ryabenki V. S., *Introduction de the theory of difference schemes*, Intersc. Publ., 1964.

[91] Gragg W. B., On extrapolation algorithms for ordinary initial value problems, *SIAM J. on Numer. Anal.*, 1965, **2**, nº 3, p. 384-403.

[92] Hartman Ph., *Ordinary differential equations*, John Wiley and sons, New York-London-Sydney, 1964.

[93] Hunter D. B., The numerical evalution of Cauchy principal values of Integrals by Romberg integrations, *Numer. Math.*, 1973, **21**, nº 3, p. 185-191.

[94] Joice D. C., Survey of extrapolation processes in numerical analysis, *SIAM Review*, 1971, **13**, nº 4, p. 435–490.

[95] Kantorovitch L. V., Krylov V. I., *Approximate methods of higher analysis*, Groningen (Nordhoff), 1958.

[96] Kantorovitch L., Akilov G., *Analyse fonctionnelle*, Ed. de Moscou, 1982.

[97] Keller H. B., Accurate difference methods for linear ordinary differential systems subject to linear constraints, *SIAM J. on Numer. Anal.*, 1969, **6**, nº 1, p. 8-30.

[98] Keller H. B., A new difference scheme for parabolic problems. In: *Numerical Solution of Partial Differential Equations*. 2. New York, Academic Press, 1971, p. 327-350.

[99] Keller H. B., Accurate difference methods for nonlinear two-point boundary value problems, *SIAM J. on Numer. Anal.*, 1974, **11**, nº 2, p. 305-320.

[100] Keller H. B., Cebeci T., Accurate numerical methods for boundary-layer flows. 2: Two-dimentional turbulent flows, *AIAA J.*, 1972, **10**, nº 9, p. 1193-1199.

[101] Kellogg B., Singularities in interface problems, Transactions of SYNSPADE, 1971, p. 351-400.

[102] Kurosh A., *Cours d'algèbre supérieure*, Ed. de Moscou, 1973.

[103] Ladyzenskaja O. A., Uralcêva N. N., *Equations elliptiques linéaires et quasi linéaires*, Dunod, 1967.

[104] Lattès R., Lions J. L., *Méthode de quasi-réversibilité et applications*, Dunod, 1967.

[105] Laurent P., Un théorème de convergence pour le procédé d'extrapolation de Richardson, *Comptes Rendus Acad. sc. Paris*, 1963, **256**, p. 1435-1437.

[106] Lavrentiev M. M., Some improperly posed problems of Mathematical Physics, Springer, Tracts in Natural Philosophy, 2 (1967).

[107] Lindberg B., On smoothing and extrapolation for the trapezoidal rule, *BIT*, 1971, **11**, p. 29-52.

[108] LINDBERG B., Error estimates and stepsize strategy for implicit midpoint rule with smoothing and extrapolation, Stockholm, Royal Inct. Of Tech., Dept. of Inf. Proc., Rept. nº A. 72.59, 1972.

[109] LINDBERG B., IMPEX — 2 — A procedure for solution of systems of stiff differential equations, Stockholm, Royal Inst. of Tech., Dept. of Inf. Proc., TRITA-NA-7303, 1973.

[110] LINDBERG B., Optimal stepsize sequences and requirements for the local error for methods for (stiff) differential equations, Univ. of Toronto, Dept. of Comput. Sci., Tech. Rept, nº 67, 1974.

[111] LIONS J., *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, 1969.

[112] MARCHOUK G., *Méthodes de calcul numérique*, Ed. de Moscou, 1980.

[113] MARCHOUK G. I., KUZNÊTSOV YU. A., *Méthodes itératives et fonctionnelles quadratiques. Sur les méthodes numériques en sciences physiques et économiques*, Dunod, 1974.

[114] MARCHOUK G. I., SHAYDOUROV V. V., A variational method for increasing the accuracy of the difference scheme, *Lecture Notes in Economics and Math. Systems*, 1976, **134**, p. 193-205.

[115] MARCHUK G. I., SHAYDOUROV V. V., Increasing of the accuracy of the projective-difference schemes, *Lecture Notes in Computer Math.*, 1974, **11**, p. 120-141.

[116] MAYERS D. F., The deferred approach to the limit in ordinary differential equations, *Computing*, 1964, **7**, p. 54-57.

[117] *Modern Numerical Methods for Ordinary Differential Equations*, Clarendon Press, Oxford, 1976.

[118] NITSCHE J., Ein Kriterium für die Quasi-Optimalität des Ritzchen Verfahrens, *Numer. Math.*, 1968, **11**, nº 4, S. 346-348.

[119] ORTENGA J. M., RHEINBOLDT W. C., *Iterative solution of nonlinear equations in several variables*, Academic Press, New York-London, 1970.

[120] PEREYRA V., On improving an approximate solution of a functional equation by deferred corrections, *Numer. Math.*, 1966, **8**, nº 4, p. 376—391.

[121] PEREYRA V., Accelerating the convergence of discretization algorithms, *SIAM J. on Numer. Anal.*, 1967, **4**, nº 4, p. 508-533.

[122] RICHARDSON L. F., The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam, *Philos. Trans. Roy. Soc., London*, ser. A, **210**, 1910, p. 307-357.

[123] RICHARDSON L. F., The deferred approach to the limit. 1 — Single lattice, *Philos. Trans. Roy. Soc., London*, ser. A, **226**, 1927, p. 299-349.

[124] RJABENKI V. S., FILIPPOW A. F., *Über die Stabilität von Differenzengleichungen*, Berlin, 1960.

[125] ROMBERG W., Vereinfachte numerische Integration, Det. Kong. Norske Videnskabers Selskab Forhandlinger, Trondheim, 1955, **28**, nº 7, p. 30-36.

[126] ROSENBROCK H. H., Some general implicit processes for the numerical solution of the differential equations, *Comp. J.*, 1963, **5**, nº 4, p. 329-330.

[127] SAMARSKI A., ANDRÊEV V., *Méthodes aux différences pour équations elliptiques*, Ed. de Moscou, 1978.

[128] SCHORTLEY G., WELLER R., The numerical solution of the Laplace's equation, *J. Appl. Phys.*, 1938, **9**, nº 5, p. 334-348.

[129] SHAMPINE L. F., WATTS H. A., DAVENPORT S. M., Solving nonstiff ordinary differential equations—the state of the art., *SIAM Review*, 1976, **18**, nº 3, p. 376-411.

[130] STETTER H. J., *Analysis of discretization methods for ordinary differential equations*, Springer-Verlag, Berlin, 1973.

[131] STOER J., BULIRSCH R., *Einführung in die numerische Mathematik*. 2. Berlin-New York, Springer-Verlag, 1973.

[132] STRANG G., FIX G. J., *An analysis of the Finite Element Methods*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1973.

[133] THATCHER R. W., The use of infinite Grid refinements at singularities in the solution of Laplace's equation, *Numer, Math.*, 1976, 25, nº 3, p. 163-178.

[134] TIKHONOV A., ARSÉNINE V., *Méthodes de résolution de problèmes mal posés*, Ed. de Moscou, 1976.

[135] VLADIMIROV V., *Distributions en physique mathématique*, Ed. de Moscou, 1979.

[136] VOÏÉVODINE V., *Algèbre linéaire*, Ed. de Moscou, 1976.

[137] WHITEMAN J. R., BARNHILL R. E., Finite element methods for elliptic mixed boundary value problems containing singularities, Proc. Conf. « Equadiff », Brno, 1972, p. 261-267.

[138] WHITTAKER E. T., WATSON G. N., *A Course of Modern Analysis*, Cambridge, Cambridge University Press, 1927.

[139] WUYTACK L., A new technique for rational extrapolation to the limit, *Numer. Math.*, 1971, 17, nº 3, p. 215-221.

[140] WYNN P., On the convergence and stability of epsilon algorithm, *SIAM J. on Numer. Anal.*, 1966, 3, nº 1, p. 91-122.

[141] YANENKO N. N., *Méthode à pas fractionnaires*, Armand Colin, 1968.

NOTATIONS

1. Domaines et frontières

2. Espacces et normes

Espaces

Normes

3. Domaines de discrétisation et leurs frontières

$\bar{\Omega}_h$ 26
$\breve{\Omega}_h$ 26
D_h 26
Ω_h 191
$\Gamma_{h,x}$ 191
$\Gamma_{h,y}$ 191
Ω'_h 191
Ω''_h 191
$\Omega'_{h,x}$ 207
$\Omega'_{h,y}$ 207
$\Omega''_{h,x}$ 207
$\Omega''_{h,y}$ 207
$\bar{Q}^\tau_h = \bar{\omega}_h \times \bar{\omega}_\tau$ 267
$Q^\tau_h = \omega_h \times \omega_\tau$ 267

$\bar{\omega}_\tau = \{t_j = j\tau,\ j = 0, 1, \ldots, M\}$ 15, 59, 266
$\breve{\omega}_\tau = \{t_{j+1/2} = (j + 1/2)\tau,\ j = 0, 1, \ldots, M - 1\}$ 16, 59
$\dot{\omega}_\tau = \{t_j = j\tau,\ j = 0, 1, \ldots, M - 1\}$ 76
$\omega_\tau = \{t_j = j\tau,\ j = 1, \ldots, M\}$ 93, 266
$\bar{\omega}_h = \{x_i = ih,\ i = 0, 1, \ldots, N\}$ 109
$\breve{\omega}_h = \{x_{i+1/2} = (i + 1/2)h,\ i = 0, 1, \ldots, N - 1\}$ 110
$\omega_h = \{x_i = ih,\ i = 1, \ldots, N - 1\}$ 110
$\bar{\omega}_h = \{0 = x_0 < x_1 < \ldots < x_N = 1\}$ réseau irrégulier 124
$\breve{\omega}_h = \{x_{i+1/2} = (x_i + x_{i+1})/2,\ i = 0, 1, \ldots, N - 1\}$ 123

4. Quotients différentiels

$u_{\mathring{t}}(t) = \dfrac{u(t + \tau/2) - u(t - \tau/2)}{\tau}$ 16

$u_{\tilde{t}}(t) = \dfrac{u(t + \tau/2) + u(t - \tau/2)}{2}$ 16

$u_t(t) = \dfrac{u(t + \tau) - u(t)}{\tau}$ 76

$u_{\bar{t}}(t) = \dfrac{u(t) - u(t - \tau)}{\tau}$ 267

$u_{\mathring{x}}(t) = \dfrac{u(x + h/2) - u(x - h/2)}{h}$ 110, 341

$u_{\tilde{x}}(x) = \dfrac{u(x + h/2) + u(x - h/2)}{2}$ 114, 341

$u_{\mathring{x}\mathring{x}}(x) = \dfrac{u(x + h) - 2u(x) + u(x - h)}{h^2}$ 343

$u_{\mathring{y}\mathring{y}}(x, y) = \dfrac{u(x, y + h) - 2u(x, y) + u(x, y - h)}{h^2}$ 192

5. Opérateurs aux différences

$D_x^{(m)}$ 196
$D_y^{(m)}$ 196
J_n^x 207
J_n^y 208

INDEX

TABLE DES MATIÈRES

À NOS LECTEURS

Les Editions Mir vous seraient très reconnaissantes de bien vouloir leur communiquer votre opinion sur le contenu de ce livre, sa traduction et sa présentation, ainsi que toute autre suggestion.

Notre adresse: Editions Mir, 2, Pervi Rijski.
péréoulok, Moscou, I-110,
GSP, U.R.S.S.

Imprimé en Roumanie